THE Philosophical Review

VOLUME 123, NUMBER 1, JANUARY 2014

Published for
the Sage School of Philosophy at Cornell University
by Duke University Press

The journal is published four times a year for the Sage School of Philosophy at Cornell University by Duke University Press. *See the inside back cover for submission guidelines.*

World Wide Web

Visit the editorial office at www.arts.cornell.edu/philrev and Duke University Press Journals at www.dukeupress.edu/journals.

Subscriptions

Direct all orders to Duke University Press, Journals Customer Service, 905 W. Main St., Suite 18B, Durham, NC 27701. Annual subscription rates: print-plus-electronic institutions, $178; print-only institutions, $168; e-only institutions, $140; individuals, $35; students, $22. For information on subscriptions to the e-Duke Journals Scholarly Collections, see www.dukeupress.edu/library/eDuke. Print subscriptions: add $14 postage and applicable HST (including 5% GST) for Canada; add $18 postage outside the US and Canada. Back volumes (institutions): $168. Single issues: institutions, $42; individuals, $12. For more information, contact Duke University Press Journals at 888-651-0122 (toll-free in the US and Canada) or 919-688-5134; subscriptions@dukeupress.edu.

Advertisements

Direct inquiries about advertising to Journals Advertising Coordinator, journals_advertising@dukeupress.edu.

Indexing and Abstracting

For a list of the sources in which the *Philosophical Review* is indexed and abstracted, see www.dukeupress.edu/philreview.

ISSN 0031-8108

MIX
Paper from
responsible sources
FSC
www.fsc.org
FSC® C012947

THE *Philosophical Review*

VOLUME 123, NUMBER 1, JANUARY 2014

CRITICAL NOTICE

BOOK REVIEWS

Regularity and Hyperreal Credences

Kenny Easwaran
University of Southern California

1. Introduction

It has been widely argued that belief is not just an all-or-nothing attitude—there is also a notion of belief that comes in degrees. Defenders of this position generally also argue that these degrees of belief, or "credences," obey something like the following principles:

- There is a set Ω of doxastic possibilities for each agent, propositions correspond to subsets of Ω, and the collection $\mathcal{F}$ of propositions in which the agent has credences is an algebra. (That is, $\mathcal{F}$ is nonempty, if a proposition is in $\mathcal{F}$, then so is its complement, and if two propositions are in $\mathcal{F}$, then so is their intersection.)[1]
- A rational agent's credences are given by a probability function P. (That is, $P(p) \geq 0$ for all propositions p, $P(\Omega) = 1$, and $P(p \cup q) = P(p) + P(q)$ whenever p and q are disjoint subsets of Ω.)

[1] Some theorists prefer to think of the objects of credence as something more sentential, rather than as sets of possibilities. The set-theoretic notation I use throughout will have to be replaced by the corresponding syntactic notation: negation in place of complement, conjunction in place of intersection, and so forth. The only significant effect this will have on my argument is that at the end of section 4, when I discuss one option that makes use of this set Ω, such a theorist will have to take the other option, which uses the conditional credence function.

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366479

- A rational agent's conditional credences satisfy the relation $P(p|q)P(q) = P(p \cap q)$.[2]

The set of doxastic possibilities represents an agent's certainties and uncertainties. Its elements may be thought of as something like possible worlds, except that they may satisfy propositions that are metaphysically impossible, or possibly even contradictory. Any proposition that the agent is not certain of must be false at some doxastic possibility. Some authors might argue that belief just is truth at all doxastic possibilities, but I suspect that many proper subsets of Ω will correspond to beliefs as well. Belief does not entail certainty, the way such a proposal would suggest.

One straightforward consequence of these principles is that if a proposition corresponds to the empty set, $\varnothing$, then a rational agent has a credence of 0 in it.[3] I will call such a proposition "doxastically impossible" because it is not true in any doxastic possibility. Many philosophers also endorse the converse:

> **Regularity:** A rational agent has credence 0 in a proposition only if it is doxastically impossible for her. Equivalently, a rational agent has credence 1 in a proposition only if it is certain for her.[4]

Some philosophers instead state a version with some sort of nondoxastic modality, especially if they think of "doxastic possibilities" as having to be logically or metaphysically possible. However, I take it that these authors are generally committed to **Regularity** as just stated, as well as some version of:

> **X-Y Transmodal Connection:** Any X'ly possible proposition is Y'ly possible.

If "Y" is interpreted as doxastic possibility for a rational agent, then this, in combination with **Regularity**, entails that rational agents are only cer-

[2] The most traditional understanding of conditional credence in fact *defines* $P(p|q)$ to equal $P(p \cap q)/P(q)$, but if we allow for the possibility that $P(q) = 0$, as is my goal in this essay, then this needs modification. I will discuss some options in sections 1 and 2, but all major proposals agree on the multiplicative formula given here.

[3] Proof: the empty set is in $\mathcal{F}$ because there is some set in $\mathcal{F}$, and the intersection of this set with its complement is the empty set. $\varnothing$ is disjoint from any proposition p, so $P(p \cup \varnothing) = P(p) + P(\varnothing)$. Since $p \cup \varnothing$ just is p, this means that $P(\varnothing) = 0$.

[4] Some authors use the term "regularity" descriptively rather than normatively. That is, for them, regularity is a property of probability functions, and there is a separate requirement for rational agents to have regular probability functions. But I will use the term "**Regularity**" to refer to the normative principle instead.

tain of logical necessities, or metaphysical necessities, or something similar.[5] Conversely, if "X" is interpreted as doxastic possibility for a rational agent, then we can interpret doxastic possibilities as possible worlds, or logical models, or something else more familiar. However, I will ignore these transmodal connections and focus instead on the principle I have called **Regularity**.

Regularity is in tension with the fact that there are specific propositions that an agent can't rule out, but for which any positive real number is clearly too high a value for the credence. I will call such propositions, as well as the numerical values (if any) of their credences, "minuscule" to avoid prejudging the question of whether their credence is 0, or if they are represented in some other way. (Arguments for the existence of minuscule propositions will be given in section 3.)

Skyrms (1980) (in a brief appendix) and Lewis (1980) (in two quick paragraphs) try to resolve this tension by suggesting that credences should not have to be real valued, but should instead be allowed to take on "infinitesimal" values as well. They point out that in the 1960s, Abraham Robinson showed the existence of mathematical structures, called "hyperreals," that behave very much like the real numbers, but include elements that are positive but smaller than any positive real number. (For instance, see Robinson 1996.) Skyrms and Lewis suggest that this theory, especially as developed by Bernstein and Wattenberg (1969), can be used to save **Regularity**, and this response has been generally accepted by philosophers working in the area for the past several decades.[6]

I think that this situation is largely based on a mistake about the role of numbers in mathematical representations. Probabilism uses a set

[5] Some authors use these principles to argue that agents should use Jeffrey's (2004) alternative to the standard update method of conditionalization since it results in one having credence 1 in one's evidence, which is generally logically and metaphysically contingent. However, if one takes standard conditionalization to produce doxastic necessity and not just credence 1, then it is compatible with the principle that I call **Regularity**, so **Regularity** itself can't be dismissed just on the trivial grounds of incompatibility with a standard update rule.

[6] As examples, see Lewis 1996, 303; Swinburne 2001, 244; Holder 2002, 296; and Norton 2007, 162. Note that Bartha and Hitchcock 1999 is *not* an instance of this sort of use of hyperreals. As they say, "we are not committed to the existence of infinitesimal degrees of belief or anything of that sort. Just as imaginary numbers can be used to facilitate the proving of theorems that exclusively concern real numbers, our use of [hyperreals] will be used to facilitate and motivate the construction of purely real-valued measures." Bartha and Hitchcock 1999, 416. However, the previously listed authors, and others following them, do suggest that agents can or must have hyperreal credences.

of possibilities to represent propositions, and real numbers to represent credences. Because the real numbers are not fine grained enough to capture all the distinctions in these doxastic states, Skyrms and Lewis argue that we should use hyperreals instead. They focus on the numerical aspect of a probabilistic representation and seek to expand it so that it can represent all the relevant distinctions. But as I presented it above, probabilism uses a *set* together with some numbers, and a *conditional* credence function as well as an unconditional one. Both of these tools describe important features of credence that shouldn't be overlooked.

In section 2, I give what I take to be the four main arguments for **Regularity** and show that these tools provide responses to all of them. In section 3, I present the problem of minuscule propositions. In section 4, I explain the hyperreals used by Lewis and Skyrms to respond to this problem, and in section 5, I show that they have too much structure to properly represent credences in ordinary propositions. Although one might think that the purely numerical representation with hyperreals is relatively simple, it turns out to have complexities far beyond those that arise from the consideration of the nonnumerical aspects of the standard representation.

This is not a definitive argument against **Regularity**, and in the appendix, I give quick overviews of a few other systems that might be used to achieve the goals that motivate it. I think pursuing probability theories based on any of these systems may be a valuable project and may help with our understanding of credence. And in fact, the hyperreals may also help, as long as we understand that they do not tell us the precise structure of credences and that not all distinctions they make should be taken to be significant. But for now, I claim that there is no reason to think credences have structure beyond that given in the opening paragraph of this introduction, with a set of doxastic possibilities, a standard real-valued probability function (which may assign 0 to doxastically possible propositions), and a standard real-valued conditional probability function.

2. Arguments for Regularity

2.1. *Learning Probability 0*

The first argument for **Regularity** is based on conditional credence. Lewis (1980, 267) says:

> I should like to assume that it makes sense to conditionalize on any but the empty proposition. Therefore I require that *C* is *regular*: *C*(*B*) is zero,

> and $C(A/B)$ is undefined, only if B is the empty proposition, true at no worlds.

The "C" Lewis refers to is the hypothetical initial credence function of a rational agent with no a posteriori information about the world. Lewis and other Bayesians suggest that the appropriate way for a rational agent to update her credences as she gains new information is to conditionalize—that is, the credence P_1 after the learning should be related to the initial credences P_0 by $P_1(A) = P_0(A|B)$, where A is any proposition in $\mathcal{F}$, and B is the proposition learned. Many philosophers follow chapter 1 of Kolmogorov 1950, where it is stipulated that $P(A|B) = \frac{P(A \cap B)}{P(B)}$, which is undefined if $P(B) = 0$. But since an agent with no a posteriori information should be able to learn *any* nonempty proposition, either every nonempty proposition must have nonzero probability (as Lewis claims), or there must be a way to update that goes beyond this standard notion of conditionalization and the standard definition of conditional probability.[7]

I will formalize the relevant version of the argument thus:

1. Any doxastically possible proposition can be learned.
2. When a rational agent learns B, she replaces her credence $P(A)$ with $P(A|B)$ for every proposition A.
3. $P(A|B)$ is defined as $\mathrm{P}(A \cap B)/P(B)$, and thus is undefined when $P(B) = 0$.
4. For a rational agent, learning can't leave all credences undefined.
5. Therefore, a rational agent doesn't have credence 0 in any doxastically possible proposition.

Premise 1 seems straightforward.[8] For premise 2, consider what Skyrms (1980, 74) says on his version of this argument:

[7] Both this and the next argument assume that updating proceeds by conditionalization. There is a commonly proposed alternative due to Richard Jeffrey, on which no single proposition needs to be learned with certainty, so that the update is compatible with maintaining uncertainty in the learned proposition, and thus is compatible with certain transmodal connections. However, this alternative still relies on $P(A|B)$, and is thus undefined if $P(A|B)$ is. Thus, using Jeffrey conditionalization instead of standard conditionalization makes no relevant difference to either of these arguments.

[8] The converse of this claim is perhaps more interesting—is it the case that everything that can be learned must be doxastically possible? It seems plausible to me that we ought to treat revisions where we give up a previous certainty as the kind that motivate an alternative to conditionalization. This situation may be more usefully studied by tech-

> How do we assimilate new knowledge of a proposition with a prior probability of zero? . . . [P]erhaps at any rate we will need external rules for some cases of belief-change not properly treated by conditionalization. But the choice should be dictated by epistemological considerations, not by the mathematics of the probability representation.

As I see it, Skyrms's point is that although premise 2 may have some problem cases, they will be epistemologically special update situations, and not the ordinary ones we normally consider, so a relevant revision of premise 2 will still leave some instances of this argument intact. Premise 4 also seems unproblematic. Thus we should focus on premise 3.

And indeed, premise 3 has serious problems. There is no need for conditional probability to be understood in terms of Kolmogorov's ratio. Many other accounts of conditional probability have been proposed that extend this account to cases where $P(B) = 0$. Perhaps the simplest modification is described in Popper 1955, according to which conditional probability is a primitive two-place function not defined in terms of unconditional probability, but freestanding. Popper's axioms guarantee that whenever $P(B) \neq 0$, the standard relations still hold, but just add the claim that $P(A|B)$ is always defined. Instead of having $P(A|B) = P(A \cap B)/P(B)$, we just have $P(A|B)P(B) = P(A \cap B)$, as I originally stated in the introduction, which can hold even when $P(B) = 0$. This sort of account is argued for by Hájek (2003), among others. Another account, quite similar to Popper's, is discussed in Rényi 1970. And in fact, although Kolmogorov (1950) stipulates the definition used in this argument in chapter 1, in chapter 5 he presents another alternative, different from the ones due to Popper and Rényi. (I argue for this account in Easwaran 2008a.) Two such accounts are compared by Seidenfeld, Schervish, and Kadane (2013).

There are a variety of options available, so there is no reason for the notion of conditional probability and its role in updating to demand **Regularity**. The ratio account is popular because it allows conditional probabilities to be defined entirely in terms of the unconditional credence function—the alternatives that I mention require in addition some sort of primitive *conditional* credence function. But Hájek argues

niques related to the AGM model of belief revision (as introduced in Alchourròn, Gärdenfors, and Makinson 1985), or one of its competitors. This gives us some epistemological considerations in favor of modifying premise 2—perhaps it ought to be prefaced with, "In any learning experience that doesn't involve giving up any doxastic certainties, . . ." But this modification is of no relevance to the argument for **Regularity**.

at length that the concept of conditional credence is at least as fundamental as that of unconditional credence, so that this mathematical parity is epistemologically significant and not just a quirk of the formalism.

Defenders of **Regularity** point out that some of these alternatives (and in particular, Popper functions) have a close connection to hyperreals, and thus suggest that they aren't really alternatives. For instance, Vann McGee (1994, 180) says,

> One approach, developed by Skyrms (1980) and Lewis (1980) is to use a nonstandard [hyperreal-valued] probability assignment in which those epistemically possible propositions that would ordinarily be assigned 0 probability are instead assigned infinitesimal probabilities. . . . The other approach, developed by Karl Popper, is more direct. . . . We shall see that these two approaches come to the same thing.

He then demonstrates that for every nonstandard-valued probability assignment, the restriction of the conditional and unconditional probability values to their "standard parts" gives a Popper function, and that every Popper function can be achieved in this way. However, the nonstandard-valued probability assignment corresponding to a given Popper function is highly nonunique—the hyperreal representation of an agent's doxastic state is far more fine-grained than the Popper-function representation corresponding to it.[9] Thus, although there is a connection between these two representations, my arguments from section 5.4 will suggest that this extra level of fine structure in the nonstandard-valued probability assignment isn't real. It can be used for a purely mathematical description of the Popper function, but one shouldn't read this extra representational power as meaning anything about the actual credences. The connection between these two options is not as tight as McGee initially claimed.

Given that there are many available accounts of conditional probability that allow for conditionalization on propositions with credence 0, for this argument to work, the defender of **Regularity** must give a non-question-begging argument in favor of analyzing conditional credence *exclusively* in terms of Kolmogorov's ratio. Absent any such argument, the most this line of reasoning can show is that there should be some way to coherently update on any doxastically possible information. If we assume

[9] In fact, there are distinct hyperreal probability assignments that correspond to the same Popper function and yet actually give rise to different decision-making behavior on the part of the agent. Halpern 2010, 168.

additionally that conditionalization is the way to update, then this tells us that the conditional credences should be well defined and should themselves form a coherent probability function. One might use the converse of premise 1, and a premise claiming that any physical, metaphysical, or logical possibility can be learned (perhaps with an exception for claims like "I do not exist"), to give an argument for some sort of transmodal connection. But none of this gives any support to **Regularity** itself.

2.2. Stubbornness

The second argument also proceeds from the rule of updating by conditionalization, but considers A rather than B in $P(A|B)$. As Lewis (1980, 268) says:

> [**Regularity**] is required as a condition of reasonableness: one who started out with an irregular credence function (and who then learned from experience by conditionalizing) would stubbornly refuse to believe some propositions no matter what the evidence in their favor.

Similarly, Skyrms (1980, 74) asks, "How can a proposition of prior probability zero come to have a posterior probability different from zero?"

As I understand the implicit argument, it starts with the mathematical fact that if $P(A) = 0$, and $P(B) \neq 0$, then $P(A|B) = 0$.[10] Thus, if an agent updates only by repeated conditionalization, and starts with $P_0(A) = 0$, then at every time t, $P_t(A) = 0$, so the agent will stubbornly refuse to believe A, no matter what the evidence. Timothy Williamson (2002, 214) gives a similar version of this argument as a reason not to accept the Bayesian picture of probability.

I will formalize the relevant argument thus:[11]

1. $P(A \cap B) = 0$ when $P(A) = 0$.
2. When an agent learns B, she replaces her credence $P(A)$ with $P(A|B)$ for every proposition A.

[10] This is a consequence of the claim that $P(A \cap B) = P(A|B)P(B)$ together with the fact that $P(A \cap B) \leq P(A)$. On the standard ratio analysis, it takes the apparently stronger form that if $P(A) = 0$, then $P(A|B) = 0$ if it exists at all.

[11] This argument makes use of the notion of "high credence." Intuitively, this should mean something like "high enough for belief," but everything about the argument, and my response to it, will work equally well if this is interpreted as "greater than 0.99999," or "greater than 0.5," or even "greater than 0.00001."

3. $P(A|B)$ is defined as $P(A \cap B)/P(B)$, and thus is 0 or undefined when $P(A \cap B) = 0$.
4. Therefore, if an agent has credence 0 in A, then she will never have high credence in A no matter what evidence B she learns.
5. For any reasonable agent, and any doxastically possible proposition A, there is some evidence B such that learning B would give the agent high credence in A.
6. Therefore, a reasonable agent does not have credence 0 in any doxastically possible proposition.

As in my previous argument, I will reject premise 3. All the proposals mentioned above on which $P(A|B)$ can be defined when $P(B) = 0$ allow it to take on any value between 0 and 1 (depending on the circumstances), even if $P(A)$ was 0. Thus, an agent can come to have high credence in A, as long as she learns some other proposition B whose initial credence was also 0.

A defender of this argument might claim that if $P(B) = 0$, then B can't be the evidence in an update. After all, most examples of minuscule propositions (to be described in section 3) involve infinite precision and may be beyond human observational capacities, so perhaps they can never constitute an agent's evidence. So if those are the only doxastically possible propositions that get probability 0, then the argument could be repaired by adding a premise that propositions with credence 0 are never learned as evidence.

But if that's right, and it's impossible for humans to learn this type of proposition as evidence, then "stubbornness" seems much less problematic—if something can never be learned as evidence, then it doesn't seem stubborn to refuse to give it high credence when learning other things. So the defender of this argument faces a dilemma: either propositions with credence 0 can be evidence, in which case premise 3 is false; or they can't, in which case stubbornness is reasonable, so premise 5 is false. Either way, one of the premises is false, so this is no sound argument for **Regularity**. (Even conceding the conclusion, one of these premises must be false.) A defender of infinitesimals might concede this point, but still object that it is strange that only propositions with credence 0 can provide enough evidence for an agent to have high credence in other propositions with credence 0. But consider the following more general argument:

1. $P(A \cap B)$ is minuscule when $P(A)$ is minuscule.
2. When an agent learns B, she replaces her credence $P(A)$ with $P(A|B)$ for every proposition A.
3. $P(A|B)$ is minuscule when $P(A \cap B)$ is minuscule and $P(B)$ is not minuscule.
4. Therefore, if A is minuscule, then the agent will never have high credence in A unless she learns some B that is also minuscule.

This argument is just a modification of the first part of the above argument, but with the notion of probability 0 generalized to the notion of being minuscule (that is, less than any positive standard real number). This argument is valid, and all the premises are accepted by Lewis, Skyrms, and other defenders of **Regularity** that appeal to hyperreals as the values of credences for minuscule propositions (and even by many defenders of alternative versions of **Regularity** that don't use Robinson's hyperreals). In particular, for premise 3 to fail, there would have to be a situation in which $P(A|B)$ and $P(B)$ are both standard positive real numbers, and yet $P(A)$ is minuscule—but this would mean that either $P(A \cap B) \neq P(A|B)P(B)$ or $P(A \cap B) > P(A)$.

Thus, defenders of hyperreals face the same issue for minuscule propositions that the opponent of **Regularity** does with probability 0. They must offer the same sort of resolution, where only minuscule propositions can provide enough evidence for one to believe other minuscule propositions. The only way to get around this is to either reject conditionalization, or revise one of the basic laws of probability for $P(A \cap B)$, either of which would destroy this argument for **Regularity**.[12]

2.3. Dutch Books

Skyrms gives a third argument for **Regularity** that is not shared by Lewis.[13] This argument extends the basic "Dutch book" argument for probabilism. The basic argument shows that if an agent's degrees of belief fail to satisfy the probability axioms, then she is vulnerable to a "Dutch book"—

[12] I thank Greg Novack and Mike Titelbaum for pressing me on this point and making me realize that I should spell out the full parallel argument for the defenders of **Regularity**.

[13] Versions of this argument were also given much earlier, in Kemeny 1955, Shimony 1955, and Stalnaker 1970, where they refer to **Regularity** as "strict coherence." I stick with Skyrms and Lewis just because they are the ones referred to by contemporary philosophers who defend **Regularity** and hyperreal credences.

a set of bets such that she considers each one individually fair or favorable (because its price is less than or equal to her degree of belief in the relevant proposition), and yet the whole set collectively guarantees her a loss. Since any rational agent views a guaranteed loss as neither fair nor favorable, then (bracketing some assumptions about evaluating a combination of bets by combining the evaluations of the individual bets) there is an inconsistency in her values.[14]

Similarly, Skyrms (1980, 74) suggests that if we allow for propositions of credence 0 to be doxastically possible, then "if we interpret probability as a fair betting quotient there is a bet which we will consider fair even though we can possibly lose it but cannot possibly win it." That is, if an agent's degree of belief in *A* is 0, then she will view as fair a bet that costs $0 with a payoff of $1 if *A* is true. However, if she is not absolutely certain that *A* is false, and she sells such a bet, then she is in a situation in which she has no possibility of making money, but a possibility of losing money, which she must surely regard as an unfavorable position, rather than a fair or favorable one.

I will formalize the argument thus:

1. Any rational agent evaluates a bet on *A* at a price equal to [her credence in *A* times the stakes] as fair to buy or sell, evaluates a bet at any lower price as favorable to buy, and evaluates a bet at any higher price as favorable to sell.
2. No rational agent evaluates a bet as fair or favorable if it gives some doxastic possibility for her to lose and no possibility to gain.
3. Selling for $0 a bet on *A* with positive stakes results in losing if *A* is true and has no possibility of gaining.
4. Therefore, no rational agent has credence 0 in any doxastically possible proposition.

The standard Dutch book argument for probabilism goes as follows:

1. Any rational agent evaluates a bet on *A* at a price equal to [her credence in *A* times the stakes] as fair to buy or sell,

[14] Some authors present the problem of vulnerability to Dutch books as a sort of pragmatic irrationality, involving the fact that an agent who is actually willing to accept each of these bets is practically irrational since she faces a guaranteed monetary loss. The interpretation I give in terms of inconsistency of values is suggested by Skyrms (1987) and Christensen (2001), and I think it is more compelling. But nothing depends on which interpretation is used.

evaluates a bet at any lower price as favorable to buy, evaluates a bet at any higher price as favorable to sell, and evaluates a combination of bets as fair or favorable if she evaluates each individual bet as fair or favorable.

2. No rational agent evaluates a combination of bets as fair or favorable if she is certain that they would collectively cause her to lose.
3. An agent's credences satisfy the probability axioms iff there is no finite collection of bets with fair or favorable prices such that she is certain they would collectively cause her to lose.
4. Therefore, a rational agent's credences satisfy the probability axioms.

There are many well-known problems involving the first premise of these Dutch book arguments (Hájek 2005, 2008). Thus, the opponent of **Regularity** could just reject this argument along with the standard Dutch book argument, by just rejecting anything resembling the first premise of either argument. But I will not take this route—I will respond to this argument in a way that is open for defenders of Dutch book arguments.

The first possibility for rejecting the argument for **Regularity** while keeping the standard Dutch book argument is to look at the difference between the second premises—in the standard Dutch book argument, there is a doxastic necessity of loss, while in the one for **Regularity**, there is only a doxastic possibility of loss. This allows room for saying that necessary loss is problematic in a way that the possible loss is not. But it seems to me that this is a bullet-biting response—it would say that a rational agent can accept a possibility of loss with no offsetting possibility of gain.

Instead, I will reject the first premise in each of these arguments and accept only a weaker premise about favorability, rather than fairness:

1′. Any rational agent evaluates a bet on *A* at any price lower than [her credence in *A* times the stakes] as favorable to buy, evaluates a bet at any higher price as favorable to sell, and evaluates a combination of bets as favorable if she evaluates each individual bet as favorable.

In this version of the premise, I have made no assumption at all about whether an agent evaluates a bet at *exactly* her credence times the stakes as fair, favorable, or unfavorable.[15] With this modification, if the first argu-

[15] In fact, some have suggested that one must evaluate bets at precisely this price as

ment is to be valid, its conclusion must be weakened to "Therefore, no rational agent has credence *less than* 0 in a doxastic possibility." This is no longer **Regularity** itself but rather a trivial consequence of the probability axioms.

However, the standard Dutch book argument can be made valid by appealing to the slightly stronger theorem that is also true:

> 3′. An agent's credences satisfy the probability axioms iff there is no finite collection of bets with favorable prices such that she is certain they would collectively cause her to lose.

For any collection of bets with fair prices such that the agent is certain they would collectively cause her to lose some positive amount, we can alter the prices by a tiny fraction of this amount, to give a collection of bets with favorable prices that have the same Dutch book property. Thus, replacing premise 1 by 1′ doesn't jeopardize the standard Dutch book arguments, so an opponent of **Regularity** can preserve the standard Dutch book argument if she is so inclined.[16]

unfavorable. Smith (1961, 5) is an early example. I will be agnostic on this point and leave open the possibility that something beyond the numerical value of one's credences is used to evaluate bets at exactly this price, so that some count as fair, some count as favorable, and some count as unfavorable. Giving a full decision theory for cases with expected value of 0 is beyond the scope of this essay.

[16] As it turns out, I think there is some motivation for defenders of countable additivity to make this modification of the argument. For any finite or infinite cardinality κ, there is a collection of κ-many bets at *fair* prices that collectively make it doxastically necessary that the agent will lose if his or her credences do not satisfy κ-additivity. Thus, with the "fair or favorable" version of the argument, we seem to get an argument for additivity of arbitrary collections of propositions. But while some probability theorists support additivity for countably infinite collections of propositions, they don't generally support additivity for uncountably infinite collections of propositions since this would rule out uniform distributions on uncountable sets, just as countable additivity rules out uniform distributions on countable sets. Thus, the defender of the "fair or favorable" version of the argument needs to either distinguish between finite and countable collections of bets (for finite additivity) or between countable and uncountable collections of bets (for standard countably-additive probabilism).

However, if we don't assume that agents will evaluate bets *exactly* at the specified price as fair, as in my modification, then we get a nonarbitrary reason to support countable additivity but not uncountable additivity. The reason there is no support for uncountable additivity is that any favorable price for selling must be positive, and the sum of uncountably many positive numbers is always infinite. Thus, selling uncountably many favorable bets on pairwise incompatible propositions never results in a guaranteed loss. But for the countable case, the Dutch book still works. If the agent buys a bet on the union of a sequence of propositions for ε less than his or her fair price and sells each bet on the nth

The defender of this argument for **Regularity** thus has to argue for the stronger premise 1 rather than the weaker 1′. One motivation would be to say that there *must* be some price at which a bet is evaluated as fair—not every price should be one that is favorable for buying or favorable for selling. But this assumption is not available to the defender of **Regularity** if the bets are monetary—money (and, by the argument I will give in section 5.4, utility too) comes only in real gradations, so any positive price for the bet is higher than the credence in a minuscule proposition and is favorable for selling, while a price of 0 is favorable for buying, and no real price is exactly fair. Only if the prices of bets themselves can have numerical values that are not standard real numbers can one maintain that every bet has a price that is exactly fair.

But this brings us to the first instance of the "numerical fallacy." When a bet has a positive real expected value, these premises say it should be evaluated as favorable. Premise 1 goes further and says that a bet with expected value 0 should be evaluated as fair. This seems plausible if we assume that the numerical expected value of a bet alone must tell us whether it is fair, favorable, or unfavorable. But if we allow that nonnumerical features of the mathematical representation of an agent's doxastic state might matter as well, then 1′ looks better. In cases where the expected value of a bet is exactly the same as the status quo, some nonnumerical feature may serve as a tiebreaker. For any proposition in which an agent has credence 0, the expected value of a bet at price 0 is exactly the same as the status quo, no matter whether the bet is bought or sold. However, the fact that in one case the agent has a possibility of winning with none of losing, and that in the other case the agent has a possibility of losing but none of winning, allows the agent to determine that one is favorable and the other is unfavorable, and neither is precisely fair.

And in fact, there are other motivations for thinking that actions might be evaluated by using some tool beyond numerical expected value. For actions with infinitely many possible outcomes, some expected values are infinite or undefined, which means that something other than numerical comparison is necessary to evaluate which are better or worse (Nover and Hájek 2004; Colyvan 2008; Easwaran 2008b). Similar issues arise if

proposition in the sequence for $\varepsilon / 2^n$ more than his or her fair price, then the total result will be exactly as if he or she had bought and sold the bets exactly at his or her fair price—which would result in a Dutch book if his or her fair prices aren't countably additive. This Dutch book parallels the one Jon Williamson (1999) gives. Exactly this point about countable versus uncountable additivity is made by Skyrms (1992, 218).

credences can be imprecise. (Adam Elga [2010] argues that there is *no* reasonable decision theory for imprecise credences, but any appropriate response to his argument will have to involve more than just single numerical expected values.) We can keep the assumption that having a greater expected value is sufficient for being preferable, but these cases already show that it is not necessary. Thus, we should reject premise 1 in both arguments and replace it by 1′. The Dutch book argument for probabilism can be saved by replacing its premise 3 by 3′. But the Dutch book argument for **Regularity** can't be saved without weakening its conclusion to a triviality.

2.4. "What 0 Means"

The final argument I will consider is rarely given explicitly, but I suspect that it is the intuitive motivation that most defenders of **Regularity** have for believing it. However, I will show that it too is an instance of the numerical fallacy. The basic idea is related to the faithfulness of mathematical representations. A statement of the idea is given in Williamson 2002, 213: "For subjective Bayesians, probability 1 is the highest possible degree of belief, which presumably is absolute certainty."[17]

The main idea is that the degree of belief function is a measure of the agent's doxastic state with respect to a proposition. This function measures propositions on a scale from 0 to 1 and assigns the value 1 to doxastic necessities and 0 to doxastic impossibilities. If the function were to assign the value 1 to some proposition other than a certainty, or 0 to some proposition other than a doxastic impossibility, then this function would not properly represent the agent's attitudes because it would falsely represent her as equally confident in two propositions that she is not

[17] Williamson follows this with a dramatic argument that an agent with such a high credence should be willing to sell for a penny a bet where the agent is tortured if the proposition comes out false. This example is related to the previous argument, but it also seems to prove too much—not only would it rule out having credence 1 in any proposition short of certainty, but it would also rule out credences of $1-\varepsilon$ for small enough ε.

Because of the well-known phenomenon of risk-aversion, it seems plausible that bets with extremely large payoffs, either positive or negative, are evaluated at least partly by some means other than expected utility. Thus, methodologically, we should limit consideration to bets with small payoffs when intuitively judging the rationality of accepting certain bets. If the possible loss is held fixed at a moderate value, while possible gain becomes extremely small, then the conclusion doesn't seem implausible. The dramatization in terms of torture is a distraction—only the claim I quoted above, about the "highest possible degree of belief," is important. I thank Lina Eriksson for this point.

equally confident in (namely, one that is doxastically contingent and one that is doxastically necessary, or impossible). Thus, if a degree of belief function properly represents an agent's attitudes, then it must satisfy **Regularity**.

I will formalize the argument as follows:

1. A doxastically possible proposition is more likely for an agent than a contradiction.
2. If p is more likely than q for a rational agent, then $P(p) > P(q)$.
3. If q is a contradiction, then $P(q) = 0$.
4. Therefore, for a rational agent, if p is doxastically possible, then $P(p) > 0$.

It is clear given my previous discussion that I will reject premise 2 as an instance of the numerical fallacy. If $P(p)$ were the complete mathematical representation of how likely p is for an agent, then this would be reasonable. But it isn't. What we need is some mathematical relation $p > q$ that says when p is more likely than q. But this relation can depend on mathematical facts beyond $P(p)$ and $P(q)$. As described in the opening paragraph of the introduction, standard probabilism gives two further mathematical features that might be relevant—the conditional probability function $P(\cdot|\cdot)$, and the set Ω of doxastic possibilities.

If we use one of the alternative accounts of conditional probability mentioned in section 2, then there will be distinctions between propositions the agent regards as certainly false, and propositions she merely has credence 0 in. For instance, on Popper's account, if $\perp$ is a contradiction, then $P(p|\perp) = 1$ for any proposition p. (In particular, $P(\neg\perp|\perp) = 1$!) It is natural to extend this behavior to other doxastic impossibilities.[18] However, Popper's axioms allow for this to fail for doxastically possible propositions whose unconditional probability is 0. Another natural picture might suggest that $P(p|q)$ is undefined when q is doxastically impossible (perhaps because an indicative-type conditional, as conditional probability is normally taken to be, makes no

[18] On a set-theoretic formulation, this is trivial because a doxastic impossibility and a contradiction are both represented by the same empty set. But even on a sentential formulation, we can prove that the behavior does extend this way if we assume that $P(p|q) = P(p|q \wedge \top)$ whenever $\top$ is a doxastic necessity. If q is a doxastic impossibility, then $\neg q$ is a doxastic necessity, so $P(p|q) = P(p|q \wedge \neg q) = P(p|\varnothing) = 1$.

sense when the antecedent is impossible), but is defined in all other situations.

On both accounts, if q is doxastically impossible, but p isn't, we will have $P(p|p \cup q) = 1$ and $P(q|p \cup q) = 0$. Thus, we can define $p > q$ as meaning that $P(p|p \cup q) > P(q|p \cup q)$ and get an ordering that validates premise 1 while falsifying premise 2. The relevant distinction can be captured in the conditional probability function rather than in the values of the unconditional probabilities. On this approach, conditional credence would turn out to be more fundamental than the $>$ relation. Such a view has been argued for by Hájek (2003). (Note that some, but not all, of the arguments there presuppose the failure of **Regularity**.)

On an approach where propositions are represented by sets of possibilities, the distinction can also be captured in the set structure of the propositions. Recall that the complete representation of the agent's credal state is the triple $(\Omega, \mathcal{F}, P)$ and not just P by itself. With this representation, we can draw the distinction between doxastically impossible propositions (which correspond to the empty set) and others (which are nonempty, even though their probability may be 0). There are many situations where it is sufficient to consider the numerical values of P and ignore the mathematical information contained in Ω and $\mathcal{F}$, but the argument under consideration only works if we assume that P is *always* sufficient. On this picture, we might say that $p > q$ iff $(P(p) > P(q)$ or $q \subset p)$. On this account, $>$ is not a total ordering, but again it validates premise 1 and falsifies premise 2.

In either case, the argument fails because premise 2 is false. Both proposals for $>$ accept the converse of premise 2 (if $P(p) > P(q)$, then $p > q$). There would be a certain elegance to accepting premise 2 as well. But it is certainly not essential to a proper mathematical theory of $>$, once one considers the nonnumerical aspects of the mathematical representation.

In fact, I will show in section 5 that the use of hyperreals to defend **Regularity** leads to problems here. Although it can save the claim that if $p > q$, then $P(p) > P(q)$, it violates the converse—there are propositions where $P(p) > P(q)$, and yet intuitively, $p \not> q$! The new numerical representation overshoots the mark, and thus equally fails to faithfully represent the agent's credences. There may be purposes for which extraneous structure is no problem, just as there may be other purposes for which some missing structure is no problem. But if a mathematical representation of credence is not to be a misrepresentation, then missing numerical structure can be made up by considering nonnumerical mathematical structure, while extra numerical structure poses a more serious problem.

However, before I can demonstrate this extra structure in the hyperreals, I must present the problem of minuscule propositions and explain the hyperreals that Lewis and Skyrms use to respond to them.

3. Minuscule Propositions

Defenders of **Regularity** have been forced to concede that some doxastically possible propositions have credence less than $1/n$ for any natural number n. It would be an interestingly bold position to deny that rational agents have credences in the propositions I will discuss in this section, or to deny that any agent may rationally treat them as doxastic possibilities, as a defender of **Regularity** without hyperreals, or another theory of infinitesimals, must do. To make clear that these propositions must have an extremely small probability (whether 0 or infinitesimal, or perhaps otherwise described), I will call such propositions "minuscule." For convenience, I will also use the term "minuscule" as a term for numbers that are less than $1/n$ for any natural number n, which are the probability values of minuscule propositions. (I will reserve the term "infinitesimal" for minuscule values that are nonzero, although some authors include 0 as an infinitesimal.)

As an example, consider a situation in which a dart is being thrown at a dart board, and consider the proposition that the center of the dart lands on the vertical line that precisely bisects the board. I claim that this proposition is a minuscule one if the agent treats the throwing of the dart as uniform, so that the probability that it lands in any given region is proportional to the area of that region.

Consider the strip around the central vertical line that is exactly $1/n$ as wide as the board is—the probability that the dart lands in this region is $1/n$, and this region entirely contains the central vertical line. Thus, the probability that the center of the dart hits the center line must be less than $1/n$ for every n. But on the other hand, it seems clear that this *could* happen, and so it seems like it should be doxastically possible—after all, nothing is special about this line to prevent the dart from hitting it, and every vertical line should be treated equally. Of course, one might worry about infinitely precise centers of darts, and the requirement that the agent distribute his credence uniformly over the board for the positions that it might hit. But as Hájek (2003) repeatedly points out, as long as a rational agent could possibly have positive credence in this setup, our account of credence should allow for it.

For another example, consider a fair coin that will be flipped infinitely many times, and consider the proposition that this coin comes up heads on every single flip. On the one hand, the probability of this proposition must be no more than $1/2^n$ for any n because that is the probability that the first n flips come up heads, which is entailed by this proposition. But on the other hand, it seems that this proposition really does describe a doxastically possible outcome of the sequence of coin flips. By the symmetry of the situation, any two sequences of coin flips should be treated similarly—there is no reason based on this setup to treat some sequences as possible and others as impossible. Of course, one might still have doubts about this situation because of the requirement that the agent believe the infinite sequence of coin flips has some possibility of actually occurring—but a denier of minuscule propositions must say that such things are not just nonactual, but doxastically impossible for every rational agent.

For a more realistic example, consider the proposition that the speed of light is exactly $2.998 \cdot 10^8$ m/s.[19] Although our measurements may have made us absolutely certain that the speed of light is not $2.997 \cdot 10^8$ m/s, or $2.999 \cdot 10^8$ m/s, there is at least some range of values that have not been ruled out by any of our experimental observations. And it seems that there is in fact some precise fact of the matter as to what this speed is.[20] However, for any n, we can surely come up with n disjoint intervals (not necessarily of equal width), such that a rational agent could regard it as equally likely (or almost equally likely) that the true value of

[19] I have been told that the speed of light actually has a stipulated value that is used as part of the definition of the meter and the second. Thus, properly speaking, I should substitute some other physical constant (like the fine-structure constant, or the exponent in some gravitational law) that has a value independent of our conventional choice of units. Further, if our theories suggest that such "constants" can actually change in value over time, then consider instead of the theory that it has a specific value, the theory that it evolves according to a particular function over time.

Maher (1990, 387–88) gives a version of this argument together with a historical claim that this accurately describes Cavendish's opinions regarding the exponent in the law of electrostatic attraction.

[20] Even if some physical theories might suggest that space-time is discrete, in a way that means there can be no such infinitely precise fact of the matter, surely we are not *completely* certain that some such theory is true. Or we can consider the beliefs of some scientist from a previous century that couldn't rationally rule out theories according to which a precise value exists.

the speed of light is somewhere in one of those intervals. Each of these intervals must have credence $1/n$, so the credence for any *particular* value contained in one of the intervals must be no greater than $1/n$. Thus, any precise specification of the value appears to be a physically realistic proposition that is doxastically possible, but for which the probability must be less than $1/n$ for any n.

Based on these three examples, and the ease of generating more like them, we should agree that there are minuscule propositions.

Defenders of **Regularity** claim that minuscule propositions must not be assigned probability 0, so if they want numerical values for the probability function, then they need some theory of infinitesimals. For Skyrms, Lewis, and their followers, Robinson's hyperreals play this role. If one instead rejects **Regularity**, one can just say that minuscule propositions have credence 0 and use objects other than the numerical probabilities, like the set of doxastic possibilities, or the conditional credence function, to represent the relevant differences.

4. What Are Robinson's Hyperreals?

In order to discuss the reasons I think that hyperreals won't be able to do the work that is demanded of them, it will be important to be clear about how they work mathematically. Skyrms and Lewis cite Bernstein and Wattenberg (1969) for a mathematically sophisticated account of how this could work, but they don't consider the details explicitly themselves. However, these details give rise to the problems I will discuss later, so I rehearse them here. (More thorough discussions are given in Luxemburg 1973 and Robinson 1996.)

Robinson's hyperreals form a mathematical structure that satisfies the complete first-order theory of the real numbers and includes a copy of the standard real numbers, together with some infinitesimals—positive elements that are smaller than any positive standard real number. Because these structures satisfy the complete first-order theory of the real numbers, much of our standard reasoning carries over to them. But it is important to note that this is only the *first-order* theory—we must be careful about statements involving *sets* of real numbers.

The proof that such structures exist is not especially complicated. It relies on a familiar result from first-order logic known as the Compactness Theorem. This result states that if Γ is a set of sentences in a first-order language, and if every finite subset of Γ has a model, then Γ has a model. There are two standard proofs of this result—importantly, both make use of nonconstructive methods, based on the Axiom of Choice. The first proof appeals to Gödel's Completeness Theorem, which states (nonconstructively) that a set of sentences has a model iff it is impossible to derive a contradiction from these sentences. Thus if Γ didn't have a model, then it would be possible to derive a contradiction from it. But since any derivation uses only finitely many sentences, this derivation would use only some finite subset Γ_0—so this finite subset Γ_0 would have no model. The second proof is preferred by model theorists, who try to avoid reference to syntactic derivations whenever possible. On this proof, the model for Γ is constructed directly from the models of its finite subsets by means of an "ultraproduct" construction, which relies on the Axiom of Choice to (nonconstructively) provide a suitable "ultrafilter." (See Chang and Keisler 1990, chapter 4, or any other model theory textbook, for details.)

Given the Compactness Theorem, Robinson's result is fairly straightforward. Let $\mathcal{L}$ be a first-order language for talking about the real numbers that includes a name for each real number, and add to it a new constant c. Let Γ be the set of all sentences in $\mathcal{L}$ that are true about the real numbers (including particular sentences like "$2 < 5$" and general ones like "$\forall x(x = 0 \vee \exists y(x \cdot y = 1))$"), together with the sentences "$c > 0$" and "$c < K$" for each K that names a positive real number. Now it is clear that every finite subset of Γ has a model—one such model will just interpret all of $\mathcal{L}$ in the standard way and interpret c as a positive real number that is smaller than any positive real number whose name is mentioned in this finite subset. But then the Compactness Theorem guarantees that Γ itself must have a model.

Because Γ includes all sentences of $\mathcal{L}$ true in the standard real numbers, the model satisfies the complete first-order theory of the real numbers. Because $\mathcal{L}$ includes names for each real number, the model includes a copy of the standard real numbers. And this model must have an interpretation for "c," which must be positive (because Γ contains "$c > 0$") and smaller than any positive standard real number (because Γ contains each $\ulcorner c < K \urcorner$). Thus, the model contains at least one infinitesimal. (Of course, c is not the only such infinitesimal—for example, $2c$, $c/5$, and c^2 will be among the infinitely many others. There will also be "infinitely large" numbers like $1/c$.)

A point that will become important later about this proof is that it is nonconstructive—both proofs of the Compactness Theorem make use of nonconstructive methods that go beyond Zermelo-Fraenkel set theory, in the completeness case to give a maximal consistent set of sentences extending a given consistent set, and in the ultraproduct case to give a nonprincipal ultrafilter to use as the base for the ultraproduct.[21] In fact, *no* constructive proof (either of the Compactness Theorem or of the existence of hyperreal structures) is possible—there is no way to exhibit a specific structure that provably shares the first-order properties of the reals and contains infinitesimals.[22]

First-order equivalence is sufficient for the basic theory of probability because it means that the standard results about addition, multiplication, and ordering still apply, including things like commutativity, associativity, existence of multiplicative inverses, and the preservation of order under multiplication or division by positive numbers. However, for some more advanced results in probability, we need second-order and higher-order expressive power, to talk about sequences, limits, and notions like topology and measurability for sets of reals.

Fortunately, even though the Compactness Theorem only applies to first-order theories, much of this higher-order work can be expressed in a first-order set theory, so that the Compactness Theorem can still be applied. One theory that will suffice is full Zermelo-Fraenkel set theory with the Axiom of Choice, but there are far weaker theories that will also suffice, such as a sort of Russellian theory of types built up off the real numbers. (See Burgess 2005 for discussions of some such systems. Section 4.4 of Chang and Keisler 1990 explicitly discusses the construction of models that include the real numbers, infinitesimals, and a theory of sets.) At any rate, we can let Γ be the set of all true first-order sentences in this much larger theory, together with sentences about some constant c that entail that it must be an infinitesimal, and the result will again be a model of this large theory that manages to include infinitesimals, while still making sense of all the constructions the original theory could talk about.

But because we are dealing only with a *first-order* theory, and not a true second-order theory, there will be some oddities with this model—for instance, not every subset of the domain will be represented by one of the objects that this model calls a "set." A true second-order theory would

[21] The Compactness Theorem and the existence of the relevant ultrafilters are both equivalent to the Boolean prime ideal theorem, which is weaker than the Axiom of Choice, but still independent of Zermelo-Fraenkel set theory (Moore 1982).

[22] Kanovei and Shelah (2004) prove that given the Axiom of Choice, there is a sentence that defines a particular hyperreal structure. But they also point out, as I will in section 5, that given ZF set theory without the Axiom of Choice, it is consistent that no hyperreal structure exists.

quantify over *all* subsets of the domain, while a first-order theory for talking about sets has a special domain of objects that play the subset role. Nothing in a first-order theory can guarantee that all subsets are represented there. Thus, there will be a distinction between the "internal" sets that the model represents with these objects and "external" sets that aren't represented by anything in the relevant model.[23] This distinction will become important later. (For more on the distinction between things that can be properly expressed in a first-order theory of sets and things that require true second-order logic, see, for instance, Boolos 1984.)

At any rate, the construction gives a model that behaves like the real numbers, includes infinitesimals, and can talk about sets and sequences. Thus, the model has the expressive power needed for probability theory. Skyrms, Lewis, and their followers hope that by using one of these models, rather than the standard real numbers, we can save **Regularity** by applying the infinitesimal values to minuscule propositions. However, the worries about external sets and sequences give some cause for concern, and I will eventually show that they doom the approach.

5. There Are No Hyperreal Credences

An important recent argument against this use of infinitesimals is Williamson 2007. In this paper, Timothy Williamson argues that infinitesimals can't be used for the case of the minuscule proposition of a fair coin coming up heads on all of its infinitely many flips. Williamson (2007, 4) says that "infinitesimal probabilities may be fine in other cases, but they do not solve the present problem." Williamson's argument rules out any such use of infinitesimals, given some weak ordering assumptions and some intuitions about the comparative probabilities of certain minuscule propositions.

I present Williamson's argument in the first subsection of this section and suggest a response in the second subsection. In the third subsection, I analyze what goes wrong with this response, and use it to show that no calculation will yield a hyperreal credence for the kind of proposition involved in this case. The fourth subsection gives the final

[23] An example of such an "external" set is the set F of all finite numbers. If this set existed in the model, then it would satisfy the following three first-order properties: $\forall x(x < 1 \rightarrow x \in F)$, $\forall x \forall y((x \in F \wedge y \in F) \rightarrow (x + y) \in F)$, $\exists x(x \notin F)$. However, in the standard model, it is clear that no such set exists. Thus, since the hyperreal model satisfies all the same first-order formulas as the standard model, it must not include such a set F.

argument that shows that credences in these propositions can't be hyperreal even if they are assigned in some noncalculational way. Unlike Williamson's argument, my argument doesn't rely on intuitions about equiprobability, but only on the supervenience of credences on the physical world. My conclusion applies only to Robinson's hyperreals, rather than other theories of infinitesimals, but it shows that hyperreals can't be the credences of *any* ordinary proposition (that is, a proposition that doesn't itself explicitly mention hyperreals, or similarly complicated mathematical objects), not just the one about infinitely many coin tosses.

5.1. Williamson's Argument

Williamson's argument proceeds as follows. Consider two fair coins that will be flipped countably many times—for definiteness, say that they will be flipped once per second, assuming that seconds from now into the future can be numbered with the natural numbers. Let the first coin be flipped starting at second 1, while the other coin is flipped starting at second 2. Let A_1 be the event that the first coin comes up heads on every single flip, A_2 be the event that the first coin comes up heads on every flip after the first, and B_1 be the event that the second coin comes up heads on every flip. By the symmetry of the situation, we might judge that $P(A_1) = P(B_1)$ because it shouldn't matter when exactly the flips occur, if they occur in the same sort of sequence. However, we might also judge that $P(B_1) = P(A_2)$ because these are corresponding sequences of flips that happen at the same moment. But $P(A_2) = 2P(A_1)$ because A_2 is independent of the first coin coming up heads on its first flip, which has probability 1/2. So $2P(A_1) = P(A_2) = P(B_1) = P(A_1)$. Subtracting $P(A_1)$ from both sides, we get that $P(A_1) = 0$. This argument works in the hyperreals because the calculation was expressed entirely in the language of first-order arithmetic.

As mentioned at the end of section 4, this is a case where the advocate of hyperreals gets too much structure. We seem to have the intuitions that $A_2 > A_1$, and yet $A_2 \not> B_1$ and $B_1 \not> A_1$. There is no way to preserve these intuitions if $>$ must correspond directly to something numerical, which presumably must give a linear ordering. No matter what values we have for $P(A_1), P(A_2), P(B_1)$, as long as $P(A_2) > P(A_1)$, it must either be the case that $P(A_2) > P(B_1)$, or (as defended by Weintraub [2008]) $P(B_1) > P(A_1)$. However, on either suggestion made at the end of section 4, both of which may allow for $>$ to be a partial ordering rather than a total ordering, the intuitions are preserved.[24] By increasing the fine-grainedness of the numerical values available, the advocate of hyperreals (or in fact *any* purely numerical representation) has made too *many* distinctions in the probability values, rather than too few. They must thus deny at least one of the intuitions in this case, in order to get $P(A_1) > 0$.

5.2. The Response?

In fact, a defender of hyperreals seems to have an argument that the probability of an infinite sequence of heads *must* be nonzero—we seem to be able to exhibit an infinitesimal that must give a lower bound on $P(A_1)$. However, it will turn out that this response proves too much and shows that *every* infinitesimal is a lower bound, so *no* value, infinitesimal or not, could possibly be the correct value. Instead of solving the problem, this attempted response makes things worse for hyperreals. But it will help demonstrate the relevance of external sets for the applications of the hyperreals, which will show that the hyperreals can't serve the purpose of aiding calculation.

The argument proceeds as a sort of dual to the argument that A_1 was a minuscule proposition. Recall that in section 3, we considered the proposition that the first n flips came up heads and showed that this proposition has probability $1/2^n$, and since this proposition is entailed by A_1, the probability of A_1 must be lower.

But imagine now that the coins will be flipped not just on every second corresponding to a natural number, but also for all the seconds corresponding to the additional infinitely large "hypernatural numbers" in some specific hyperreal structure.[25] (Ignore for the moment that these

[24] Defining $p \succ q$ iff ($P(p) > P(q)$ or $q \subset p$), this works because $P(A_1) = P(B_1) = P(A_2) = 0$ and $A_1 \subsetneq A_2$, while B_1 is neither a subset nor a superset of either of the other two events. Defining $p \succ q$ iff $P(p|p \cup q) > P(q|p \cup q)$, we have to be a bit more careful. As long as $P(A_1|A_1 \cup B_1)$ and $P(A_2|A_2 \cup B_1)$ are both undefined, this suggestion will work as well.

Williamson claims that we can't have $A_1 \succ \varnothing$, but his argument assumes that $B_1 \succeq A_2$. Both of the models just given show that if B_1 and A_2 can be incomparable, rather than equally likely, then it can be the case that $A_1 \succ \varnothing$. Williamson claims that we have an intuition that B_1 is equiprobable with A_2, but I claim that our intuition is just that B_1 is neither more nor less probable than A_2 and that we can't reliably tell the difference between these types of intuition. At any rate, the subtle differences between equiprobability and incomparable probabilities with the same numerical value (Williamson and I agree that these events all have probability 0) make such intuition-based arguments more difficult.

[25] To show that these infinitely large natural numbers must exist, recall that the

additional flips change the case and may thus change the relevant probabilities.) Now consider the claim that the coin comes up heads on every flip up to some hypernatural N, and not just on the flips corresponding to standard natural numbers. This proposition entails that every flip in the original infinite sequence comes up heads (since the sequence up to N includes all the standard natural numbers and more), and thus $P(A_1)$ must be at least as large as its probability. But the probability of this claim seems to be $1/2^N$, which is a nonzero infinitesimal. Thus, it appears that $P(A_1)$ must be larger than some infinitesimal, and not equal to 0 as Williamson's argument suggested!

However, this argument turns out to be too powerful. Let ε be any positive infinitesimal hyperreal. Then a version of this argument will show that $P(A_1) > \varepsilon$. Since ε is infinitesimal, $1/\varepsilon$ is larger than every natural number. For any real number $x > 0$, there is an integer power of 2 between x and $x/2$. Since this is a first-order claim, the nonstandard model must satisfy it as well—when x is $1/\varepsilon$, call the relevant number 2^N. N must be a hypernatural number since otherwise 2^{N+1} would be a standard natural number larger than $1/\varepsilon$. But now consider the claim that every flip up to N comes up heads. This proposition still entails A_1, but it has probability $1/2^N$, which is greater than ε.

Thus, we see that $P(A_1) > \varepsilon$, as claimed. Since this is the case for *every* ε, this means that although any positive real number is too large to be $P(A_1)$, every infinitesimal is too small—but by definition, there is nothing smaller than every positive real number except for these infinitesimals. So *no* value is possible.

5.3. Calculations with Internal and External Sets

The problem with these arguments is that we are trying to use a *nonstandard* model to calculate the probability that every *standard* flip comes up heads. If we are using a nonstandard model that can talk about sets of numbers as well as numbers, then it turns out that the set of all standard natural numbers is an "external" set that this model can't talk about—

standard model satisfies the claim that for every x there is a natural number between x and $x + 1$, and also the claim that every real number has a multiplicative inverse. Since these are first-order claims, the nonstandard model must satisfy them as well. If ε is some infinitesimal, then $1/\varepsilon$ must be infinitely large—ε is less than $1/n$ for every standard natural number, so $1/\varepsilon$ must be larger than each n. But any "natural number" N between $1/\varepsilon$ and $1/\varepsilon + 1$ must be an infinitely large natural number, which we can call a "hypernatural number."

therefore, it should be no surprise that this model can't be used to calculate a specific probability for events defined in terms of this set.

To show that the set of standard natural numbers is external, consider the normal argument that the probability that the first N flips all come up heads is $1/2^N$. This argument works by induction. If $N = 1$, then the probability that the first N flips all come up heads is clearly $1/2^1$. Now, we assume that the claim is true for N and show that it is true for $N + 1$. The next flip of the coin is fair, and thus has probability $1/2$ of coming up heads. The first N flips and the next flip are independent, and so the probability that the first N flips come up heads *and* the next one does is the product of their two probabilities, which is $1/2^N \cdot 1/2 = 1/2^{N+1}$. Thus, by induction, this must be true for all N.

But induction is a second-order principle. It says that for any *set* of natural numbers, if the set contains 1, and contains $N + 1$ whenever it contains N, then the set contains all positive natural numbers. But notice that in a hyperreal model, the set of *standard* natural numbers violates this principle since it leaves out the hypernaturals. If the language and logic used for calculations with infinitesimals (and other nonstandard numbers) has an induction principle that holds for all sets that it recognizes, then the set of standard natural numbers is not a set internal to this model, so it can't tell us anything about the probability of an event essentially involving the set of standard natural numbers, like the one Williamson is interested in. Conversely, if the model does give a way to calculate the probability of this event, then it doesn't satisfy the full induction principle, and there is no way to calculate the probability of N flips all coming up heads. Either way, the attempted response to Williamson's argument fails.

And this holds more generally, not just in the example that Williamson considers. If we use the hyperreals to calculate the probability of a proposition, then there are three possibilities. The proposition might be an "ordinary" proposition, which the language can express using only standard first-order vocabulary (such as the proposition that the first 739 flips come up heads, or that the dart falls exactly on the center line of the board). The proposition might be one that the language can express, but only using vocabulary that refers to particular nonstandard elements of the hyperreal model (such as the proposition that the first N flips come up heads, or that the dart falls within $1/N$ of the center line of the board, where N is a particular hypernatural number). Or the proposition might be one that the language can't express at all (like the proposition that every standard flip comes up heads).

In the first case, since the calculation is first-order and uses only standard vocabulary, the first-order equivalence between the hyperreals and the standard reals means that the result must be the same as if we calculated with the standard model—so the result can't be infinitesimal. In the third case, we just can't use the model to do the calculation—we need some extended technique. Only in the second case can this method assign an infinitesimal value. But these cases can't provide an argument for the use of hyperreals in describing mental states since they already presuppose that propositions involving hyperreals get credences. At any rate, the *ordinary* minuscule propositions discussed in section 3 must get probability 0. And this would mean giving up **Regularity**, which was a primary motivation for using the hyperreals in the first place.

To sum up: the argument against Williamson's assignment of probability 0 to an infinite sequence of heads failed because it tried to do a calculation on a set external to the language. And this is a general problem for the hyperreals—any proposition expressible in standard vocabulary whose probability is calculated in a hyperreal model must get a standard probability.[26]

5.4. *The Complexity Argument*

In response to these earlier points, a defender of **Regularity** might suggest that hyperreal probabilities are assigned in some language-external way that doesn't involve any calculation within the model.[27] In this section, I will show that this sort of response can't work—at least, any such assignment of hyperreal values to credences in ordinary propositions (ones that can be expressed using only standard vocabulary) will impute some structure that actual credences of physical agents themselves can't have.

Although Bayesianism concerns itself with idealized rational agents, and not the imperfect physical beings we encounter in our daily life, I claim that the essentially nonphysical nature of agents with hyperreal credences makes them irrelevant for the epistemology of physical agents. The other idealizations, of logical omniscience and the like, are

[26] A similar argument against the possibility of infinitesimal *chances* rather than credences is given in Barrett 2010.

[27] In effect, this is how the proposal in Bernstein and Wattenberg 1969 works, which is cited as a model by both Skyrms and Lewis. The *hyperreal* interval [0,1] is broken up into N segments, where N is some particular infinitely large hypernatural number, and this division is used to assign probabilities to various subsets of the *standard* interval [0,1] (without hyperreals) so that every singleton has nonzero probability.

not physically impossible, and we can make sense of a way in which actual imperfect agents might become more and more like these idealized agents.[28] These idealizations are like the ones from physics involving frictionless surfaces, and infinitely deep water for waves to travel on. But where these idealizations involve the *removal* of some limitation, the hyperreals involve the *addition* of nonphysical structure. Although I phrase my argument in terms of the actual credences of physical agents, it works just as well for any rational requirement on physical agents. Just as no agent could have a credence that was a particular hyperreal, no agent could have a rational requirement involving some particular hyperreal.

The premises and conclusion of the argument are as follows:

1. Credences supervene on the physical, in the sense that there is a function that takes as input a complete mathematical description of the physical world, and a specification of an agent and a proposition, and returns as output the number representing the credence of the agent in that proposition.[29]
2. The function relating credences to the physical is not so complex that its existence is independent of Zermelo-Fraenkel set theory (ZF).
3. All physical quantities can be entirely parameterized using the standard real numbers.
4. The existence of a function with standard real number inputs and hyperreal outputs is independent of ZF.
5. Therefore, credences in ordinary propositions (ones expressible without mention of hyperreals or closely related notions) do not have hyperreal values.

[28] In fact, the statement of Bayesianism from the first paragraph of the introduction doesn't even involve this much idealization. Because of the use of doxastic possibilities, there may be logical necessities that the agent fails to have credence 1 in. Because there is no diachronic rule of updating, there is no requirement of perfect memory. There may still be some sort of idealization involved in the construction of the set of doxastic possibilities, but we can think of this set as being in a way implicitly defined by the entirety of the physical facts about the agent, even though no particular doxastic possibility is represented by any particular thing in the agent's brain.

[29] In the sections defending **Regularity**, I was very interested in the nonnumerical aspects of credence, but the discussion here of hyperreal credences is just about the numerical representation.

The first two premises of the argument express a form of physical supervenience about credences—there couldn't be two worlds that agree on the entirety of the physical facts and yet are different in terms of the credence a particular agent has in a particular proposition, and the pattern of dependence is (in some very generalized sense) computable. Premise 3 is an assumption about the structure of the actual physical world. Together, these first three premises entail (given only standard set theory) that there is a function that takes a standard real number description of the universe as an input, together with a specification of an agent and a proposition, and gives that agent's credence in that proposition as an output. Premise 4 is a mathematical result that I will demonstrate further on, and it implies that this function can't take a proposition described entirely in terms of standard real numbers and give a hyperreal output, which is the conclusion of the argument.

While premises 1 and 3 might be controversial, it is only essential to my argument that they be at least plausible. The defender of hyperreal credences must deny at least one of these assumptions, which would entail doing serious physics, or philosophy of mind. It seems wrong to judge the answers to these questions based on an epistemological principle like **Regularity**. One should have independent grounds for rejecting these assumptions in order to reject my conclusion. (But see note 31 for a further concern about rejecting premise 3.)

Premise 2 can be motivated as a version of the Church-Turing thesis. This thesis states that all intuitively computable functions can be computed by Turing machines. Many authors have suggested stronger versions saying that in fact any mental process whatsoever can be simulated by a Turing machine. They have often defended this claim by appeal to an even stronger principle stating that any *physical* process can be simulated by a Turing machine. Since anything simulated by a Turing machine can be proven to exist within the framework of ZF set theory, without appeal to anything more complicated, this would entail premise 2. And of course, premise 2 is much weaker—there are plenty of noncomputable functions that can be perfectly well defined within ZF (for instance, Turing's original "halting function," and most other standard examples of noncomputable functions). Of course, the strong physical version of the Church-Turing thesis may be implausible, as argued by Copeland and Sylvan (1999) (as well as by many others). But proposed challenges to it only go a few levels up the Turing hierarchy, and don't come anywhere near the complexity level of ZF, much less beyond it. There's no clear motivation for thinking that the interpretation of physical processes as mental ones should introduce this particular type of logical complexity, unless one were already committed to using hyperreals or something similar.

My argument doesn't make any assumptions about what form the physical realization of credences takes. If an agent's mental state must include a concrete representation within her brain of every single proposition that she has credences in, together with a representation of the value of that credence, then I might be able to strengthen the conclusion to show that *no* proposition gets hyperreal credence. The defender of **Regularity** might use this sort of picture to argue that physical agents can't have credences in the sorts of infinitary propositions argued to be minuscule. But on most accounts, mental states can involve physical processes outside the agent's brain and can be dispositional in ways that don't require explicit representation of every proposition or doxastic possibility.

My assumptions are consistent with the following scenario. Perhaps an agent can have dispositional credences just by having a commitment to some kind of uniformity over her doxastic possibilities. The agent might be unsure whether a particular dartboard with width one meter is properly parameterized by the real numbers or by the hyperreals, and be committed to credence 1/2 in each.[30] Her commitment to uniformity may be sufficient to fix her conditional credence in every proposition of the form "the exact center of the dart hits some point within *x* meters of the vertical line at the center of the board" to be 2*x*, conditional on the board being parameterized by the hyperreals. If so, then for any particular hyperreal *x*, the agent will dispositionally have hyperreal credence in this proposition, even though he or she is unable to grasp the proposition directly. Of course, such a proposition is not an "ordinary" proposition since we need to use a hyperreal to even state it. But my argument shows that even on such a dispositional account of credences, physical agents don't have hyperreal credences in ordinary propositions.

Something like premise 3 is clearly essential for an argument like this to work. If the physical world really does involve magnitudes with the structure of the hyperreals, then it is not hard to see how agents might conceivably have hyperreal credences.[31] For instance, it

[30] Premise 3 entails that in fact every dartboard is properly parameterized by the real numbers. But, as already mentioned, this fact is compatible with at least some reasonable agents being unsure of it.

[31] Interestingly, although hyperreal physics might allow hyperreal credences, it may

could be that credences in particular propositions are given by the precise voltage drop across some particular neuron or synapse in the agent's brain. If voltages can be hyperreal, then these sorts of credences can be too.[32] But my argument shows that if none of the fundamental physical quantities have hyperreal structure, then even a substantially more complicated realization of credences (possibly involving not just the voltage across a particular synapse, but states of the entire network of neurons, or causal connections to the external world, or even a radical version of the extended mind hypothesis [Clark and Chalmers 1998]) can't give rise to hyperreal structure in the credences. This is why premise 1 appeals to a description of the full physical world and a specification of the agent, rather than just a physical description of the agent.

Now I will argue for premise 4. (This argument is given in the first footnote of Kanovei and Shelah 2004.) There are various results due to Robert Solovay and Hugh Woodin showing that, assuming the existence of certain large cardinals, it is consistent with ZF (without the Axiom of Choice) that there are no nonmeasurable sets of real numbers (Neeman 2010). However, given a nonstandard hyperreal number, one can define a nonmeasurable set of real numbers.[33] Thus, it is compatible with ZF set theory that there are no functions that give a nonstandard hyperreal output for any standard real-valued inputs. However, ZF together with the Axiom of Choice does prove the existence of such functions. Thus, the existence of such functions is independent of ZF, which (by premise 2) means that they are too complex to properly represent the physical

not suffice to save **Regularity**. If we consider the dartboard example again, then we can see that an agent's credence that the center of the dart hits *precisely* the center line of the dartboard will have to be even smaller than any of the infinitesimals available from the hyperreal structure used in physics. So we will need credences to have some even finer hyperreal structure than the physics. And I suspect that a variant of this overall argument will rule out this sort of mismatch between the physical hyperreals and the ones used for credences.

[32] If chances are themselves fundamental physical quantities, rather than themselves being realized by other fundamental physical quantities, then the existence of hyperreal chances (as argued for by Hofweber [forthcoming]) could be enough for there to be hyperreal credences. But as in footnote 4, this may not save **Regularity**.

[33] One version of this proof is in Luxemburg 1973, 66–67. Another version is given by Terence Tao at terrytao.wordpress.com/2008/10/14/non-measurable-sets-via-non-standard-analysis/. The construction involved is actually very similar to the Bernstein and Wattenberg construction of hyperreal probabilities—although the construction gives every singleton a nonzero probability, it also shows that some more complex sets can't get *any* probability, real or hyperreal.

manifestation of credences. Thus, the credences of physical agents in ordinary propositions are not hyperreal.

This argument is in many ways just a sharpening of the argument given in section 5 of Hájek 2003. Hájek makes the argument that infinitesimal probability assignments are "defective" because they are "ineffable." That is, we have no way to pick out which infinitesimal is the one assigned to any given proposition. My claim is the more specific one that no physical facts could make one of these infinitesimals rather than another be the credences of a particular agent. Although the Axiom of Choice guarantees that such hyperreal-valued functions exist, and although these functions are quite useful to talk about in mathematical contexts, they have mathematical structure that goes beyond that of credences.

None of this rules out a certain instrumental use of hyperreals. For instance, as mentioned in note 6, Bartha and Hitchcock (1999) use hyperreals to describe a particular standard real-valued probability function. In many cases, it may be more convenient for a theorist to describe credences by using a hyperreal-valued function than to use the set of possibilities Ω, the algebra of propositions $\mathcal{F}$, the probability function P, and a conditional probability function. But the structure of the hyperreals goes beyond the physical structure of credences, while $(\Omega, \mathcal{F}, P)$ doesn't seem to. Thus, if we want our mathematical theory to faithfully represent the structure of credences, as supposed by the arguments for **Regularity**, then we should prefer the nonnumerical structure of the standard representation over the apparent convenience of the numerical structure of the hyperreals.

6. Conclusion

I have shown that the arguments in favor of **Regularity** are all unsound. The mathematical structure of probability theory (especially when supplemented with a conditional probability function) involves several features that can do the work that nonzero values are supposed to do in these arguments. There is no need for betting behavior or comparative probability to be represented purely by individual numbers in the system. Additionally, the particular numbers endorsed by Skyrms, Lewis, and their followers (namely, the Robinson-style hyperreals) have too much mathematical structure to represent anything about any physically possible agents. The advantage that the hyperreals have is that they are first-order equivalent to the standard reals. However, they are so unlike the standard reals in terms of second-order logic (with the distinction between internal and external sets) and set-theoretic complexity that they can't provide a faithful model of credences of the sort wanted by defenders of **Regularity**. There are of course many other number systems that are simpler than the hyperreals, which may be promising for this purpose, and I canvass several such systems in the appendix. But the basic point still stands—the mathematical structure surrounding the standard real number representation of credence appears to provide an adequate representation of credences, despite giving up **Regularity**. Any extension of this system that is intended to save **Regularity** should avoid introducing extra complexities like those of the hyperreals.

A. Appendix: Alternative Theories of Infinitesimals

Although this essay argues that Robinson-style hyperreals can't be the values of credences, there are several other frameworks that have been proposed that can reasonably be called "infinitesimal probabilities." It may be that some of these systems do a better job of representing the epistemic structure of credences than the version of the Kolmogorov picture that I defend in the main text, and so they merit further study. But the question of whether they save **Regularity** comes down to the question of whether these are considered to be standard real numbers with further mathematical structure, or whether the structure as a whole constitutes a new number system. It seems to me that this is a relatively empty terminological question, and thus research on these versions of the theory should focus on the extent to which they do or don't respect the epistemology, and not on whether or not they happen to assign a "number" that looks like 0 to a doxastically possible proposition.

A.1. Carnap

Carnap was already aware of the problem of minuscule propositions in 1960, before Robinson's construction of the hyperreals.[34] Since there was at that time no known rigorous mathematics of infinitesimals, Carnap sought to outline what such a theory ought to look like, in order for infinitesimals to do the work he wanted for probability. The draft he wrote eventually appeared posthumously as Carnap 1980.

[34] I thank Branden Fitelson for pointing out to me the papers discussed in this section.

In this draft, Carnap posed four problems whose solution would yield a notion of infinitesimal probabilities, together with partial solutions to the first and third problems. The first problem is to lay down axioms that can be used to characterize the relations explicating the notion of one set of real numbers being smaller than another, and one set being *infinitely* smaller than another. He proposed about twenty conditions that these two relations should jointly satisfy and listed some theorems that follow from them. The second problem was to give an explicit characterization of some relation on sets of real numbers that would satisfy these axioms, which he was unable to do.

The third problem is to investigate the equivalence classes of sets of real numbers under the "same size" relation characterized in the first two problems. He carried out this project to the extent of showing that these equivalence classes could be considered as themselves constituting a number system that contained infinitesimals. The fourth problem is to give an explicit characterization of this number system and a function assigning values from this number system to sets of real numbers.

Given his partial characterizations, Carnap was able to give some characterization of what the number system might look like. In particular, just as in Robinson's later system, there would be some infinite set of infinitesimal numbers ε_i, such that for any two of them, one would be infinitely smaller than the other. However, unlike in Robinson's system, smaller infinitesimals would be "absorbed" into larger ones, so that if ε_1 is infinitely smaller than ε_2, then $\varepsilon_1 + \varepsilon_2 = \varepsilon_2$ and $\varepsilon_1/\varepsilon_2 = 0$, which in Robinson's system would happen only if $\varepsilon_1 = 0$. This might have interesting implications for the relation between conditional and unconditional probability.

In the same volume where this draft was first published, Douglas Hoover (1980) published a short note using Robinson's system (in particular, with the construction given by Parikh and Parnes [1974]) to shed some light on Carnap's problems. In particular, he showed that the Parikh and Parnes system satisfies all but two of Carnap's axioms and that those two axioms were inconsistent with the others in any case, so that nothing better could be hoped for. However, the resulting number system is the Robinson-style hyperreals, which (as mentioned above) behave somewhat differently from the number system Carnap envisioned. A similar account that also uses the hyperreals is given by Benci, Horsten, and Wenmackers (2013). But perhaps some other system satisfies those two axioms while rejecting some others and behaves more like the system Carnap hoped for.

A.2. Lexicographic Probabilities

Another approach to minuscule propositions is the technique of "lexicographic probabilities." Versions of this approach appeared as early as Kemeny 1955. Although Kemeny generally imposes the requirement of "strict coherence" (his term for **Regularity**), on pages 270–72, he considers what happens if this requirement is removed.

In his system, probabilities are assigned to sentences from some finite language. He shows that if strict coherence is required, then the probability functions are determined by assignments of nonzero numbers to the state descriptions (maximal consistent conjunctions of atomic sentences and their negations), summing to 1. The conditional probability $P(a|b)$ (Kemeny uses the notation "$P(a, b)$") is then given by the sum of the values on the state descriptions that make both a and b true, divided by the sum of the values on the state descriptions that make b true.

If strict coherence is not required, the situation is a bit more interesting. Instead of a single assignment of numbers to the state descriptions, we need a *sequence* of such assignments, such that the values in each individual assignment add up to 1, and such that every state description gets a nonzero value on exactly one of the assignments. In this case, the conditional probability $P(a|b)$ is defined as before, except that the values used in the calculation are the values given by the *first* assignment in the sequence where some state description making b true has a nonzero value.

This construction has since been generalized by others, including van Fraassen (1995) and Halpern (2010). In the modern version, we consider an arbitrary well-ordered sequence of probability functions, with the requirement that every nonempty proposition get a nonzero value in some function in the sequence. To update on a proposition, one first removes from the sequence all functions that give this proposition the value 0, and then applies standard conditionalization to all remaining functions. At any point in time, only the first function in the sequence represent's the agent's credences, with the others only serving to encode information about conditional credences and updates.

Van Fraassen and Halpern both consider the relation between these lexicographic probabilities and Popper's functions with primitive conditional probabilities. As suggested by Kemeny, Halpern shows that there is a strong equivalence between Popper functions and these lexicographic probabilities if we impose particular relations between the assignments. That is, the countably additive lexicographic probabilities and the countably additive Popper functions can be paired up in such a way that corresponding lexicographic probabilities and Popper functions give exactly the same conditional probabilities to every pair of propositions. However, Halpern also shows (in examples 3.2 and 3.5) that if we impose a somewhat stronger relation between different assignments in the sequence, or give up on countable additivity, then there are Popper functions to which no lexicographic probability corresponds.

Both authors also consider the relation between lexicographic probabilities and probability functions that are allowed to take hyperreal values. Van Fraassen, in appendix A4, cites McGee's (1994) result showing that there is a correspondence between Popper functions and hyperreal probabilities and goes on to argue that the Popper functions (or associated lexicographic probabilities) are superior to the hyperreal-valued functions. In particular, he cites the fact that hyperreal values are highly nonunique and that the lexicographic probabilities are much easier to construct given a sequence of conditional probability values that one wants to match.

Halpern shows that the relations between these three approaches are somewhat more subtle. Although taking the "standard part" (the real number closest to a given hyperreal) of every conditional probability in a hyperreal-valued probability function gives a Popper function, and every Popper function arises in this way, Halpern shows that on a natural way of interpreting decision theory in these two frameworks, the corresponding functions give rise to different preferences among gambles. As Halpern shows in his example 5.3, an agent who has credence $1/2 + \varepsilon$ in p and $1/2 - \varepsilon$ in $\neg p$ will prefer a payoff conditional on p to the *same* payoff conditional on $\neg p$, but will *disprefer* it to any *larger* payoff conditional on $\neg p$. Since there is no Popper function with this behavior, the Popper function corresponding to this hyperreal-valued probability function fails to properly represent it.

As it turns out, the correspondence between lexicographic probabilities and hyperreal-valued probability functions doesn't have this problem—but as mentioned above, in infinite probability spaces where countable additivity isn't required, the correspondence is only one way. There are hyperreal-valued probability functions that are not represented by any lexicographic probability.

Thus, Popper functions, lexicographic probabilities, and hyperreal-valued probability functions are very similar in behavior (much more similar than Carnap's proposal is to any of them), but there are still important differences. In particular, hyperreal values allow far more fine-grained distinctions than either of these other options. Additionally, since Popper functions and lexicographic probabilities are both definable in very constructive ways, the arguments I give in section 5 against hyperreals don't cause problems for Popper functions or lexicographic probabilities.

Whether these lexicographic probabilities really represent "infinitesimal credences" or count as a way to satisfy **Regularity** seem to be primarily terminological questions. We can say that a proposition is minuscule if it gets the value 0 from the first function in the sequence and say that its credence is "infinitesimal" if it gets a nonzero value from some later function in the sequence. But we might also just identify credences with the value assigned by the first function in the sequence, which would interpret these lexicographic probabilities as violating **Regularity**.

A.3. Further Mathematical Theories of Infinitesimals That Could Be Applied

There are also some other mathematical theories of infinitesimals that could be used in place of Robinson's hyperreals. For instance, one could use the theory of "surreal numbers" developed by John Conway, or the techniques of "smooth infinitesimal analysis" based on the ideas of William Lawvere. (See Conway 1976 for the former and Bell 1998 for the latter.) Smooth infinitesimal analysis doesn't seem like an especially promising formalism since it treats infinitesimals as more like "infinitely small line segments" rather than as points on a number line, and it requires intuitionist logic instead of classical logic. The surreal numbers seem more promising as a device for future philosophers of probability to use. Their construction is a simultaneous generalization of Dedekind's construction of the real numbers and von Neumann's construction of the ordinals and can be carried out in a very weak set theory. As it turns out, we can name particular surreal infinitesimals, like $1/\omega$ and $2^{-\omega}$. However, the use of surreal numbers for probability values will have to be substantially different from the way Skyrms and Lewis recommend using hyperreals because the technique they take from Bernstein and Wattenberg (1969) leads directly to the construction of nonmeasurable sets, and thus goes beyond ZF in some substantial way. Additionally, the defender of surreal probabilities will need to address the worries raised by Williamson's argument about linearly ordered comparative probabilities. It would be interesting to see whether the use of surreal numbers could get around these worries, but the eventual theory will have to look substantially different from the one that Skyrms and Lewis proposed and other philosophers have uncritically adopted.

References

Alchourròn, C. E., P. Gärdenfors, and D. Makinson. 1985. "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *Journal of Symbolic Logic* 50: 510–30.

Barrett, M. 2010. "The Possibility of Infinitesimal Chances." In *The Place of Probability in Science*, ed. E. Eells and J. H. Fetzer, Boston Studies in the Philosophy of Science, 65–79. Dordrecht: Springer.

Bartha, P., and C. Hitchcock. 1999. "The Shooting-Room Paradox and Conditionalizing on Measurably Challenged Sets." *Synthese* 118: 403–37.

Bell, J. L. 1998. *A Primer of Infinitesimal Analysis.* New York: Cambridge University Press.

Benci, V., L. Horsten, and S. Wenmackers. 2013. "Non-Archimedean Probability." *Milan Journal of Mathematics* 81, no. 1: 121–51.

Bernstein, A. R., and F. Wattenberg. 1969. "Non-standard Measure Theory." In *Applications of Model Theory of Algebra, Analysis, and Probability*, ed. W. A. J. Luxemburg, 171–86. New York: Holt, Rinehart and Winston.

Boolos, G. 1984. "To Be Is to Be a Value of a Variable (or to Be Some Values of Some Variables)." *Journal of Philosophy* 81, no. 8: 430–49.

Burgess, J. 2005. *Fixing Frege.* Princeton: Princeton University Press.

Carnap, R. 1980. "The Problem of a More General Concept of Regularity." In *Studies in Inductive Logic and Probability*, ed. R. Jeffrey, 2:145–55. Berkeley: University of California Press.

Chang, C. C., and H. J. Keisler. 1990. *Model Theory.* New York: Elsevier.

Christensen, D. 2001. "Preference-Based Arguments for Probabilism." *Philosophy of Science* 68, no. 3: 356–76.

Clark, A., and D. Chalmers. 1998. "The Extended Mind." *Analysis* 58: 10–23.

Colyvan, M. 2008. "Relative Expectation Theory." *Journal of Philosophy* 105, no. 1: 37–44.

Conway, J. 1976. *On Numbers and Games.* Natick, MA: A K Peters.

Copeland, B. J., and R. Sylvan. 1999. "Beyond the Universal Turing Machine." *Australasian Journal of Philosophy* 77, no. 1: 46–66.

Easwaran, K. 2008a. "The Foundations of Conditional Probability." PhD diss., University of California, Berkeley.

———. 2008b. "Strong and Weak Expectations." *Mind* 117, no. 467: 633–41.

Elga, A. 2010. "Subjective Probabilities Should Be Sharp." *Philosophers' Imprint* 10, no. 5: 1–11.

Hájek, A. 2003. "What Conditional Probability Could Not Be." *Synthese* 137: 273–323.

———. 2005. "Scotching Dutch Books?" *Philosophical Perspectives* 19: 139–51.

———. 2008. "Arguments for—or against—Probabilism." *British Journal for the Philosophy of Science* 59, no. 4: 793–819.

Halpern, J. 2010. "Lexicographic Probability, Conditional Probability, and Nonstandard Probability." *Games and Economic Behavior* 68, no. 1: 155–79.

Hofweber, T. Forthcoming. "Infinitesimal Chances." *Philosophers Imprint.*

Holder, R. 2002. "Fine-Tuning, Multiple Universes and Theism." *Noûs* 36, no. 2: 295–312.

Hoover, D. 1980. "A Note on Regularity." In *Studies in Inductive Logic and Probability,* ed. R. Jeffrey, 2:295–297. Berkeley: University of California Press.

Jeffrey, R. 2004. *Subjective Probability: The Real Thing.* New York: Cambridge University Press.

Kanovei, V., and S. Shelah. 2004. "A Definable Nonstandard Model of the Reals." *Journal of Symbolic Logic* 69, no. 1: 159–64.

Kemeny, J. 1955. "Fair Bets and Inductive Probabilities." *Journal of Symbolic Logic* 20, no. 3: 263–73.

Kolmogorov, A. N. 1950. *Foundations of the Theory of Probability.* New York: Chelsea.

Lewis, D. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability,* ed. R. Jeffrey, 263–93. Berkeley: University of California Press.

———. 1996. "Desire as Belief 2." *Mind* 105, no. 418: 303–13.

Luxemburg, W. A. 1973. "What Is Nonstandard Analysis?" *American Mathematical Monthly* 80, no. 6: 38–67.

Maher, P. 1990. "Acceptance without Belief." In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association,* 1:381–92. Chicago: University of Chicago Press.

McGee, V. 1994. "Learning the Impossible." In *Probability and Conditionals,* ed. E. Eells and B. Skyrms, 179–99. New York: Cambridge University Press.

Moore, G. 1982. *Zermelo's Axiom of Choice: Its Origins, Development, and Influence.* Mineola, NY: Springer-Verlag.

Neeman, I. 2010. "Determinacy in $L(\mathbb{R})$." In *Handbook of Set Theory,* ed. M. Foreman and A. Kanamori, 1887–1950. Dordrecht: Springer.

Norton, J. 2007. "Probability Disassembled." *British Journal for the Philosophy of Science* 58: 141–71.

Nover, H., and A. Hájek. 2004. "Vexing Expectations." *Mind* 113: 305–17.

Parikh, R., and M. Parnes. 1974. "Conditional Probabilities and Uniform Sets." In *Victoria Symposium on Nonstandard Analysis,* ed. A. Hurd and P. Loeb, 180–94. Berlin: Springer-Verlag.

Popper, K. 1955. "Two Autonomous Axiom Systems for the Calculus of Probabilities." *British Journal for the Philosophy of Science* 6, no. 1: 51–57.

Rényi, A. 1970. *Foundations of Probability.* San Francisco: Holden-Day.

Robinson, A. 1996. *Non-standard Analysis.* Princeton: Princeton University Press.

Seidenfeld, Teddy, Mark J. Schervish, and Joseph B. Kadane. 2013. "Two Theories of Conditional Probability and Non-Conglomerability." Eighth International Symposium on Imprecise Probability: Theory and Applications, July 2–5, Compiègne, France.

Shimony, A. 1955. "Coherence and the Axioms of Confirmation." *Journal of Symbolic Logic* 20, no. 1: 1–28.

Skyrms, B. 1980. *Causal Necessity.* New Haven: Yale University Press.

———. 1987. "Coherence." In *Scientific Inquiry in Philosophical Perspective,* 225–42. Pittsburgh, PA: University of Pittsburgh Press.

———. 1992. "Coherence, Probability and Induction." *Philosophical Issues* 2: 215–26.

Smith, C. 1961. "Consistency in Statistical Inference and Decision." *Journal of the Royal Statistical Society, Series B (Methodological)* 23, no. 1: 1–37.

Stalnaker, R. 1970. "Probability and Conditionals." *Philosophy of Science* 37, no. 1: 64–80.

Swinburne, R. 2001. *Epistemic Justification.* Oxford: Oxford University Press.

van Fraassen, B. 1995. "Fine-Grained Opinion, Probability, and the Logic of Full Belief." *Journal of Philosophical Logic* 24: 349–77.

Weintraub, R. 2008. "How Probable Is an Infinite Sequence of Heads? A Reply to Williamson." *Analysis* 68, no. 3: 247–50.

Williamson, J. 1999. "Countable Additivity and Subjective Probability." *British Journal for the Philosophy of Science* 50: 401–16.

Williamson, T. 2002. *Knowledge and Its Limits.* Oxford: Oxford University Press.

———. 2007. "How Probable Is an Infinite Sequence of Heads?" *Analysis* 67, no. 3: 173–80.

Does Division Multiply Desert?

Theron Pummer
University of California, San Diego

1. Dividing a Killer

Consider the following hypothetical case:

> ***Killer.*** Last week I killed an innocent old lady. I did this because she unwittingly cut in front of me at the grocery store. I committed *murder.* I possess free will in whatever sense is necessary to ground the claim that I deserve punishment, in some retributive sense, for my act.[1] During the past week I thought deeply about what I did, realized that it was very wrong, and have truly turned over a new leaf. I once was vicious, but now am virtuous.[2] Punishing me would benefit no one. For example, it would have no deterrence effects.

Although not everyone agrees, many believe that I ought to be punished for what I did last week. This particular belief is entailed by a more general belief many have:

> ***Desert.*** When people culpably do very wrong or bad acts, they deserve punishment in the following sense: at least other things being equal,

I am very grateful to Dick Arneson, David Brink, Tim Campbell, Roger Crisp, Mark Johnston, R. J. Leland, Jeff McMahan, Per Milam, Dana Nelkin, Derek Parfit, Morgan Parker, Sam Rickless, Melinda Roberts, Nate Rockwood, Danny Weltman, Mason Westfall, the editors and an anonymous referee at the *Philosophical Review,* and audience members at the 2012 Rocky Mountain Ethics Congress and the 2012 Berkeley-Stanford-Davis Graduate Philosophy Conference for their extremely helpful comments on this essay.

[1] For a powerful case against desert-grounding free will, see Pereboom 2001. I will here assume, for the sake of argument, that we do possess such free will.

[2] By "now am virtuous," I mean that my behavior now is at least not worthy of any punishment.

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366487

> they ought to be made worse off, simply in virtue of the fact that they culpably did wrong—even if they have repented, are now virtuous, and punishing them would benefit no one.

In this essay, I will discuss some issues concerning Desert that arise in cases where people divide.[3] Here is such a case:

> ***Killer's Division***. A week after I killed the old lady in *Killer*, I got on my bicycle, and I headed home from the soup kitchen—where I started volunteering as part of turning over a new leaf. A drunk driver slammed into me, completely destroying my legs and torso, and cracking my skull open on the pavement. The impact disconnected the left and right hemispheres of my brain. Luckily, it is the technologically advanced future, and these two parts of my brain were immediately scooped up off the pavement and rushed to the hospital. The left hemisphere of my brain was transplanted into a cloned body just like my previous one. The right hemisphere of my brain was transplanted into a different cloned body also just like my previous one.
>
> Had *only* the left hemisphere of my brain survived and been successfully transplanted into a cloned body, *I* would have survived. And, owing to sufficient redundancies in my brain, I would have continued on with just this one hemisphere exactly as I would have if I had not been in the accident.[4] The analogous counterfactual, pertaining to the right hemisphere of my brain, is true. But both hemispheres did survive.
>
> The person who inherited the left hemisphere is called *Lefty*, and the person who inherited the right hemisphere is (predictably) called *Righty*. Note that "Lefty" refers to "the person who inherited the left hemisphere," *whether or not this person is identical to me* ("Righty" is being used analogously).
>
> I went unconscious the moment I was hit by the drunk driver. The hospital, knowing that I own two cottages, one hundred miles apart, sent Lefty home to one cottage and Righty to the other. Each woke up the next morning in their respective beds completely unaware of what happened after the drunk driver began to suspiciously swerve toward the bicycle.

We can now ask some crucial questions: *Does Lefty deserve to be punished for what I did last week? Does Righty?* Some people think that the answers to these questions depend on whether or not Lefty or Righty are the same person as me. They accept

[3] I am referring here to the sort of division cases famously explored by Derek Parfit (1984, chap. 12).

[4] Such redundancies, though not realistic, are metaphysically possible.

> ***Desert Requires Identity.*** In order for one to deserve punishment for some act, one must be the same person as the person who performed this act.[5]

If this view were true, it would be very important whether or not, for example, Lefty is me. Now, I cannot be the same person as Lefty *and* be the same person as Righty since if this were true, then Lefty and Righty would be the same person (by the transitivity of identity). But since Lefty and Righty wake up in separate beds, tickling Lefty would cause him but not Righty to laugh, and so forth, they are separate persons (by the indiscernibility of identicals). So, I am either one but not the other, or I am neither. Which is it? We might accept a

> ***Reductionist View.*** The fact of personal identity is reducible to psychological or physical facts.

But now notice that whatever physical or psychological facts we could point to that would make it the case that Lefty is me would equally make it the case that Righty is me. Thus, if it is psychological or physical facts that would make it the case that I am identical to Lefty or to Righty (as Reductionist Views imply), then it seems implausible that I could be one but not the other, and would thus have to be neither. But we might instead accept a

> ***Nonreductionist View.*** The fact of personal identity is not reducible to psychological or physical facts.

If this view were true, it might be that I *am* Lefty, even though I am physically and psychologically related to Righty in every way that I am physically and psychologically related to Lefty. (Similarly, it might be that I *am* Righty, and so forth.) For reasons I cannot rehearse here, many are inclined, with Parfit, to reject Nonreductionist Views. However, I should clarify that the potential puzzles for Desert I here discuss do not arise only if we reject Nonreductionist Views, or only if we accept Parfit's theory of personal identity. They arise for a very wide and heterogeneous class of theories of personal identity.

But they do not arise for *all* theories of personal identity. They do not arise according to theories that imply that *Killer's Division*, as I have described it, is impossible. Two important counterfactual claims are

[5] John Locke espouses Desert Requires Identity. Interpretive evidence for this can be found from sec. 13 through sec. 26 of "Of Identity and Diversity" in Locke 1975 [1694].

included in the description of *Killer's Division*. They are claims about who I would be if either of the following cases occurred:

> ***Only Lefty Survives.*** The right hemisphere of my brain was destroyed in the accident. Only Lefty survives.
>
> ***Only Righty Survives.*** The left hemisphere of my brain was destroyed in the accident. Only Righty survives.

The two claims are

(i) that in *Only Lefty Survives*, Lefty and I would be the same person; and,
(ii) that in *Only Righty Survives*, Righty and I would be the same person.

Whenever both (i) and (ii) are true, Lefty and Righty are what I will call my *continuers*. (Recall that, for example, "Lefty" refers to "the person who inherited the left hemisphere," whether or not this person is identical to me.)

Some theories of personal identity imply that (i) or (ii) is false, and they are thus inconsistent with my description of *Killer's Division*. If such theories are false, such that division is possible, then there are some potential puzzles for Desert. And I will here assume *arguendo* that such theories are false.

Again, there is substantial variety within the wide class of theories of personal identity according to which these Desert puzzles *do* arise. There are different sorts of Reductionist Views. *Psychological Views* say that personal identity is reducible to certain psychological facts.[6] And according to the standard such view, X at time t_n is the same person as Y at later time t_m if and only if Y is uniquely psychologically continuous with X.[7] *Physical Views* say that personal identity is reducible to certain physical facts.[8] According to a standard such view, X at time t_n is the same individual as Y at later time t_m if and only if Y is uniquely physically con-

[6] Defenses of such views appear in Locke 1975 [1694], S. Shoemaker 1970 and 1984, Parfit 1971 and 1984, Perry 1972, Lewis 1976, Nozick 1981, and elsewhere.

[7] I here follow the formulation offered by David Shoemaker (2009, 61), which is a standard formulation, compatible with those offered by many prominent defenders of Psychological Views (including Derek Parfit and Sydney Shoemaker).

[8] For examples of such views, see Thomson 1997, Olson 1997, and DeGrazia 2005.

tinuous with X.[9] Personal division is clearly possible on nearly all Psychological Views, and on many but not all Physical Views.[10] Division is also possible according to many but not all Nonreductionist Views.[11]

[9] I write "same individual as" rather than "same person as" because on many Physical Views, particular persons are numerically identical to particular nonpersons, for example, presentient fetuses. On many such views, persons are not *essentially* persons, but rather are essentially biological organisms, or brains.

[10] Division is possible on many Physical Views. They imply that, in *Killer's Division,* both Lefty and Righty are physically continuous with me in the same way that my future self would be physically continuous with my past self. In *Only Lefty Survives,* Lefty would be uniquely physically continuous with me. In *Only Righty Survives,* Righty would be uniquely physically continuous with me. These Physical Views thus imply (i) and (ii), the two crucial counterfactual claims included in the description of *Killer's Division.*

There are, however, *some* Physical Views that imply that (i) or (ii) is false. For example, consider the claim that, in order for Y to be relevantly physically continuous with X, Y must possess at least roughly half of X's brainstem. Since even today's best neurologists cannot successfully divide brainstems, it is highly dubious that both Lefty and Righty would each possess at least roughly half of my brainstem. Since at least one of them would not possess enough of my brainstem, at least one of them would fail to be relevantly physically continuous with me. And so either (i) or (ii) would be false, depending on whether it is Lefty or Righty who gets (most of) my brainstem. Thus, what I described happening in *Killer's Division* would in fact not happen.

In response, one could argue directly against the claim that in order for Y to be relevantly physically continuous with X, Y must possess at least roughly half of X's brainstem. Indeed, it seems irrelevant to my survival that I maintain one and the same brainstem. An exact replica of my brainstem, which performed the same basic regulatory functions just as well as my original brainstem, would also seem to preserve my existence just as well. See D. Shoemaker 2009, 106.

Alternatively, one could point out that, while brainstems cannot be divided by today's best neurologists, perhaps brainstems will be divisible by *tomorrow's* best neurologists. At the very least, one could point out that it is not *metaphysically impossible* for my brainstem to divide, such that Lefty and Righty would each get half—and that exact replicas of the half each is missing would immediately regrow from each original half. And to generate puzzles for Desert, *Killer's Division* need not be more than metaphysically possible.

But for all that, there remain some Physical Views that imply that *Killer's Division* is metaphysically impossible. For example, some Physical Views might have stricter requirements on physical continuity, and thus on identity. These requirements might imply that if a person loses half of his or her brainstem, he or she ceases to exist—even if an exact replica of the lost half were immediately regrown from the remaining half. Insofar as they imply that division is impossible, I will here assume *arguendo* that such Physical Views are false.

[11] See, for example, Chisholm 1976, Swinburne 1984, and Merricks 1998 for defenses of Nonreductionist Views.

There are at least two different sorts of Nonreductionist View. According to *Soul Views,* X at time t_n is the same person as Y at later time t_m if and only if X and Y possess the same soul (or Cartesian Ego). According to *Simple Views,* X at time t_n is the same person as Y at later time t_m if and only if X and Y are the same person.

There are some Reductionist Views that characterize division differently than I have so far. I have been assuming that if division is possible, and if Reductionism is true (such that it is implausible that I am Lefty but not Righty, or Righty but not Lefty), then there are *three* distinct persons in *Killer's Division*: me, Lefty, and Righty. But David Lewis (1976) argues that there are only *two* persons: *me-Lefty* and *me-Righty*. Prior to the division, me-Lefty and me-Righty exist simultaneously and are colocated. They go their separate ways at the point of division.

While there are some minor implications of shifting to Lewis's metaphysics,[12] doing so does not yield any significantly different implications for the main issues about Desert here discussed. The same issues will arise, but under somewhat different presentations.

In sum, the assumption that division is possible is a very ecumenical one, and the theories according to which division is possible form a very wide and heterogeneous class. It thus seems well worth exploring the puzzles to which such theories might give rise. But I grant that insofar as my puzzles trouble Desert-believers at all, they are unlikely to trouble those who antecedently believed division to be impossible. And some Desert-believers might even argue that my puzzles provide reason to

Division might be possible, according Soul Views, if souls could split into halves, and if a person could survive with at least half of his or her soul. Then Lefty and Righty could each inherit half of my soul, and (i) and (ii), the two crucial counterfactual claims included in the description of *Killer's Division*, would be true.

But perhaps souls cannot split. Division, as I have described it, is still possible on some nonsplitting Soul Views. Here is how: If only Lefty survived, he would get my soul. If only Righty survived, he would get my soul. Thus, (i) and (ii) are true. If both Lefty and Righty survived, then only one of them would get my soul (only God knows which one), and the other would get a different soul. Similarly, division is possible on some Simple Views. Here is how: If only Lefty survived, he would be me. If only Righty survived, he would be me. Thus, (i) and (ii) are true. If both Lefty and Righty survived, then only one of them would be me. (In all such cases, Lefty and Righty would be my continuers, whether or not they are me.)

However, division is impossible on some nonsplitting Soul Views and on some Simple Views. A nonsplitting Soul View might say that my soul always goes with Lefty. If Lefty does not exist, then neither do I. Then (ii) would be false. And a Simple View might say that the fact about my identity always goes with Righty. If Righty does not exist, then neither do I. Then (i) would be false. Or such views might say or imply that *either* (i) *or* (ii) is false, without committing to *which* is false. See Kagan 2012a, 150–62, for a very clear discussion of division, which contains some interesting remarks about division and Soul Views.

[12] If Lewis's view is correct, then the conflict between Desert Requires Identity and Irrelevance of Others (which is noted below in section 3) disappears.

believe that division is impossible—though, for reasons that I cannot explicate here, I think we should be reluctant to accept such arguments.

Having introduced *Killer's Division*, and having made these preliminary remarks about personal identity and the possibility of division, I can now introduce the first potential puzzle.[13]

2. The Multiplication Argument

In addition to Desert, consider two further claims:

> ***Irrelevance of Division.*** The total amount of punishment that is deserved cannot increase merely in virtue of personal division.
>
> ***Irrelevance of Others.*** How much punishment a person deserves cannot be affected by the mere existence or nonexistence of another person. (For the technically more accurate articulation of this claim, see the footnote.)[14]

These two claims might seem, considered independently, hard to deny. However, together they threaten to undermine Desert, as the following ***Multiplication Argument*** shows. Throughout this Multiplication Argument, please read "P deserves punishment" as "P deserves *X amount* of punishment."

> (1) In *Killer*, I deserve punishment for what I did last week. (Desert)
>
> (2) If, in *Killer*, I deserve punishment for what I did last week, then in *Only Lefty Survives*, Lefty deserves punishment for what I did last week.

[13] In section 5, I will begin to explore some cases where persons *fuse*. What I said in this section about the possibility of division applies, mutatis mutandis, to the possibility of fusion. That is, I am here assuming that fusion is possible (and that theories of personal identity that imply that fusion is impossible are false).

[14] Purely for convenience, I have decided to use somewhat loose language in spelling out Irrelevance of Others (similar to the way Parfit [1984, 267] formulated Williams's first requirement). But it is important to note that I am interpreting Irrelevance of Others to imply that Lefty—the person with the left hemisphere—cannot deserve more or less punishment depending on the existence or nonexistence of others, whether or not he is identical to me, and whether or not Lefty in *Only Lefty Survives* is identical to Lefty in *Killer's Division*. Put more precisely, the idea is that if in one possible world w the person with the left hemisphere deserves X punishment in virtue of his relation to process p, then in any world w^* in which the person with the left hemisphere is related to process p in intrinsically exactly the way the person with the left hemisphere is in w, in w^* the person with the left hemisphere deserves X punishment (inspired by Johnston 1989, 381). While it might have to be read twice, this technically more accurate articulation of Irrelevance of Others remains intuitively plausible.

So, (3) In *Only Lefty Survives*, Lefty deserves punishment for what I did last week. (1 & 2)

So, (4) In *Killer's Division*, Lefty deserves punishment for what I did last week. (3 & Irrelevance of Others)

(5) If, in *Killer*, I deserve punishment for what I did last week, then in *Only Righty Survives*, Righty deserves punishment for what I did last week.

So, (6) In *Only Righty Survives*, Righty deserves punishment for what I did last week. (1 & 5)

So, (7) In *Killer's Division*, Righty deserves punishment for what I did last week. (6 & Irrelevance of Others)

So, (8) In *Killer's Division*, both Lefty and Righty deserve punishment for what I did last week. (4 & 7)

(9) If, in *Killer's Division*, both Lefty and Righty deserve punishment for what I did last week, then the total amount of punishment that is deserved can increase merely in virtue of personal division. (If each deserves X, the total deserved is 2X.)[15]

So, (10) The total amount of punishment that is deserved can increase merely in virtue of personal division. (8 & 9)

Contradiction. (10 & Irrelevance of Division)

If (2) through (10) and Irrelevance of Division are true, then the claim that in *Killer* I deserve punishment for what I did last week (1, entailed by Desert) must be false. On the other hand, if (1) through (10) are true, then Irrelevance of Division must be false, and it must be the case that

> ***Division Multiplies Desert.*** When a person who deserves punishment undergoes division, each product of division deserves the same amount of punishment this person deserves.

If I deserve X amount of punishment in *Killer*, then Division Multiplies Desert implies that Lefty and Righty, the products of my division, each deserve X amount of punishment in *Killer's Division*. The total amount of punishment deserved has thus increased from X to 2X. Hence the name of this argument: the *Multiplication* Argument.

[15] Remember that, for example, (8) should be read as "In *Killer's Division*, both Lefty and Righty deserve *X amount of* punishment for what I did last week."

3. Is the Existence of Others Relevant to Desert?

One possible response to the Multiplication Argument is to deny Irrelevance of Others. But before exploring this response, it is important to observe—as Parfit famously did—that identity does not matter for rational prudential concern. I should be just as prudentially concerned about what happens to Lefty as I should be prudentially concerned about what happens to my future self, whether or not I am identical to Lefty (same goes for me and Righty).[16] Suppose Lefty's quality of life in *Only Lefty Survives* would be the same as in *Killer's Division.* If I knew my cerebral hemispheres were about to split, I would have *no* prudential reason whatsoever to take a pill that would cause my right hemisphere to liquefy upon becoming disconnected from my left hemisphere, thereby ensuring that only Lefty would survive. Why mention this? If we believed that identity matters for rational prudential concern, then we might believe that, insofar as I do not exist in *Killer's Division,* I have already gotten at least some of what I deserve. What I have gotten, we might claim, is as good as the death penalty. Or we might more modestly claim that division, while not as bad as death, still isn't as good as ordinary survival. Either claim might move us to reject (4) and (7). However, since both claims are implausible, we cannot plausibly reject (4) or (7) on such grounds. What happens to me in *Killer's Division* is at least as good as ordinary survival. Having made this preliminary observation, I will now consider some further objections to the Multiplication Argument.

Irrelevance of Others licenses the move from the claims that Lefty (3) and Righty (6) deserve punishment in a case in which only one of them survives to the respective claims that Lefty (4) and Righty (7) deserve punishment in a case in which they both survive. If Lefty deserves punishment in one case, then Lefty also deserves punishment in another case that is exactly the same except that some other person exists.

But people who accept Desert Requires Identity would reject Irrelevance of Others. They would claim that the *Only Lefty Survives* and *Only Righty Survives* cases are importantly different from *Killer's Division* because (on some views) facts about personal identity change between the former cases and the latter case. Moreover, for those who accept Desert Requires Identity, it matters, in the latter kind of case, whether a Nonreductionist View is true.

[16] And Lefty should be just as prudentially concerned about what *happened* to me, whether or not he is identical to me (same goes for Righty and me).

First, suppose Nonreductionist Views are false (some Reductionist View is true). Then, as explained above, I am *neither* Lefty nor Righty in *Killer's Division*. Defenders of Desert Requires Identity will thus claim that neither Lefty nor Righty deserves punishment for what *I*, a separate person, did last week. They will thereby deny (4) and (7).

Next, suppose that a Nonreductionist View is true. Now, as explained above, it is possible that I *am* Lefty, even though I am physically and psychologically related to Righty in every way that I am physically and psychologically related to Lefty. (Similarly, it is possible that I *am* Righty, and so on.) Defenders of Desert Requires Identity will here claim that Lefty *or* Righty, *but not both*, deserve punishment for what I did last week. They will thereby deny either (4) or (7), depending on whether I am Lefty or Righty.[17]

However, Desert Requires Identity does not seem plausible in division cases. In nondivision cases, the following italicized question seems to garner intuitive support for Desert Requires Identity: *how can* **I** *deserve punishment for what* **someone else** *did?* But in *Killer's Division*, Lefty is my *continuer.* That is, while he is not identical to me,[18] he would have been had it not been for Righty's existence. It is only a technicality involving the logic of identity that prevents Lefty from being me.[19] Now imagine Lefty asking: *how can* **I** *deserve punishment for what* **someone else** *did?* If the "someone else" Lefty is referring to here is me, then the answer to his question seems easy: *because* **you** *are this person's continuer.* We do not believe that Lefty can get off scot-free owing to a technicality involving the logic of identity. While metaphysical facts about personal identity might be contingent on whether, say, only Lefty survives or both Lefty and Righty survive, it seems deeply implausible that something as serious and important as whether someone deserves punishment for committing a murder could be. And while the logic of identity might *force* us to accept that I am Lefty in *Only Lefty Survives* but not in *Killer's Division*, it cannot analogously force us to accept that Lefty would deserve more or less punishment, depending on whether Righty survives. Desert Re-

[17] Of course, in such cases we might not *know* the fact of personal identity, or who has my soul (or Cartesian Ego), and so, according to Desert Requires Identity, not know whether it is Lefty or Righty who deserves punishment.

[18] If a Nonreductionist View were true, we could here assume that the further fact of personal identity holds between me and Righty.

[19] The technicality is that Lefty and Righty are not identical (indiscernibility of identicals), and so I cannot both be identical to Lefty and be identical to Righty (since the transitivity of identity would then imply that Lefty and Righty are identical).

quires Identity does not provide a plausible answer to the Multiplication Argument.[20]

It is important to notice that, in denying Desert Requires Identity, I am only claiming that if Desert is true, then my continuers might well deserve punishment for my wrongdoing, even if they are not identical to me. I am not committing to any more specific view about what it is about my continuers that makes them deserving of punishment for what I did. I am not, for example, committing to a view that says that L deserves to be punished for T's wrongdoing if and only if L is psychologically continuous with T.

We might deny (4) and (7) for a different reason. Recall that (4) says that, in *Killer's Division*, Lefty deserves X amount of punishment for what I did last week and that (7) is the analogous claim about Righty. Again, the moves from (3) and (6) to (4) and (7), respectively, are licensed by Irrelevance of Others. However, we might deny Irrelevance of Others, and instead accept

> ***Divided Desert.*** When a person who deserves punishment undergoes division, each product of division deserves an equal proper fraction of the total amount of punishment this person deserves. These products must inherit *equal* fractions because they are alike in all relevant respects.[21]

Divided Desert implies that if I deserve X amount of punishment in *Killer*, Lefty and Righty each deserve X/2 amount of punishment in *Killer's*

[20] We could maintain both Desert Requires Identity and Irrelevance of Others if we denied Desert. Moreover, perhaps staunch believers in Desert Requires Identity would regard my division cases, combined with Irrelevance of Others, as an argument against Desert. I suspect, however, that most people attracted to Desert would, upon encountering my division cases, willingly abandon Desert Requires Identity and regard its plausibility as limited to ordinary, nondivision cases.

[21] Parfit (1984, 271–72) asks, "If the malefactor is sentenced to twenty years in prison, should each resulting person [from division] serve twenty years, or only ten?" The latter disjunct suggests Divided Desert. Later on, in chapters 14 and 15, Parfit considers a variety of extreme and moderate implications of his views on personal identity. Here is a sketch of his argument for the extreme claim concerning Desert (from Parfit 1984, 324, and 1986, 838–39):

P1	Desert requires Nonreductionist identity.
P2	There is no Nonreductionist identity.
So, C	Desert is false.

I am not aware of anyone (including Parfit) who has been persuaded by this argument to abandon Desert. I suspect that this is because people who are attracted to Desert and who accept a Reductionist View would have no qualms about denying P1.

Division. But this violation of Irrelevance of Others also seems implausible. It cannot be that, through the sheer luck that Righty survived, Lefty would deserve less punishment.[22]

One might offer the following counterargument: It *is* true that how much punishment Lefty deserves depends on the existence of another person—namely, me. And so we have a counterexample to Irrelevance of Others. And so we cannot plausibly invoke Irrelevance of Others in response to Divided Desert.

This counterargument fails. First, what Irrelevance of Others implies is that the *mere* existence or nonexistence of another person cannot affect how much Lefty deserves. And it is not my mere existence that would make Lefty deserving of punishment, but that I killed an innocent old lady last week, and that Lefty is my continuer. Second, it is perhaps helpful to see that, while I am neither Lefty nor Righty, and while Lefty and Righty are nonidentical, certain prudential and moral relations hold between me and Lefty and between me and Righty that do not hold between Lefty and Righty. Whereas Lefty and Righty are my continuers, Lefty is not Righty's continuer, and Righty is not Lefty's continuer. Accordingly, I should have prudential concern both for Lefty and for Righty (the way I ordinarily would for my future self), but Lefty and Righty should not have such prudential concern for each other (though perhaps they should have special concern for each other in something like the way siblings or close friends do).[23] Similarly, while the wrongs I did can matter for how much punishment Lefty and Righty deserve, the wrongs that Lefty does cannot matter for how much punishment Righty deserves, and so on.[24]

[22] Luck can certainly affect how much punishment a person *can* or *will* suffer, but not how much he or she *deserves* to suffer. Satan would not be less deserving of punishment if he found and escaped into a bunker that made him invulnerable to punishment.

[23] Distinguishing between the predivision and postdivision individuals in this way can, I believe, solve one of Parfit's puzzles about the *Branch-Line Case* (see Parfit 1984, 287–89), but I cannot get into this here. Also see part 1 of Velleman 2008.

[24] We might consider

> *Divide and Rob*. I know that I am about to divide, and I form two intentions: the intention to have Lefty rob a bank, and the intention to have Righty write a check to Against Malaria Foundation. I divide. Lefty robs the bank on the basis of the first intention I formed. Righty writes the check on the basis of the second intention I formed.

Some might be tempted to claim that the wrongdoing that Lefty did *does*, in this case, matter for how much punishment Righty deserves. But we should be careful not to misidentify what it is in virtue of which Righty might deserve punishment. If Righty deserved

It seems that the mere existence or nonexistence of Righty (Lefty) could not affect how much punishment Lefty (Righty) deserves. Attempts to deny (4) and (7) by denying Irrelevance of Others seem too implausible.

4. Does Division Multiply Desert?

Suppose we simply accepted (1) through (10) of the Multiplication Argument. This would imply Division Multiplies Desert and that Irrelevance of Division is false. If Division Multiplies Desert, then, for example, if I deserve twenty years of punishment in *Killer,* Lefty and Righty would each deserve twenty years in *Killer's Division*—making a total of forty years. And we might not see what is so implausible about this result. In this section, I will mention a couple of implications of Division Multiplies Desert, which some might find hard to believe.

First, it might seem that if anything increases how much punishment is deserved, it is increases in things of the following sort: the severity or number of bad acts or motives, the degree to which persons are virtuous or vicious, and the degree to which the relevant people are culpable for these acts, motives, or characters. More generally, some might believe we should accept the

> ***Fault Restriction.*** There cannot be a greater total amount of deserved punishment if there is no increase in fault. ("Fault" is here construed, rather broadly, as any kind of error for which an agent is relevantly culpable.)

And indeed personal division does not per se involve any increase in fault, as I here understand it. If I were a vicious person just prior to my division, which consequently resulted in two vicious people, then my division would arguably involve an increase in overall fault. But in *Killer* I am, prior to my division, a virtuous person;[25] Lefty and Righty would thus

punishment in *Divide and Rob,* it might be in virtue of his being a continuer of someone who formed the bad intention on the basis of which Lefty acted in robbing the bank. Lefty's crime might then be relevant in that it could serve as *evidence* of a bad intention that I, the person of whom Righty is a continuer, had. Moreover, Righty might deserve punishment if he knew that Lefty was going to rob a bank, but did nothing to stop him, or to warn the bank, police, and so forth. But it seems implausible that Righty could deserve more or less punishment *merely* in virtue of whether or not Lefty in fact commits the bank robbery.

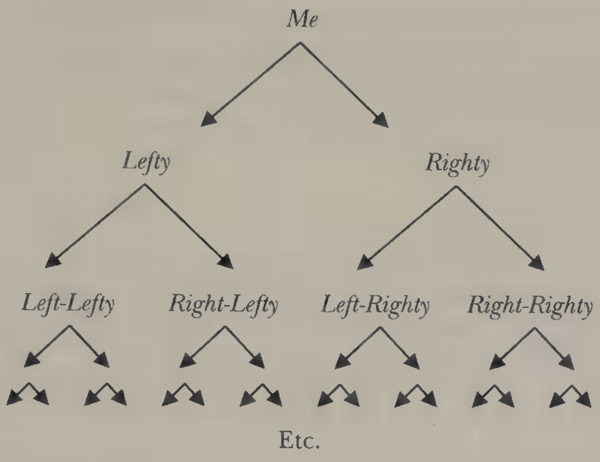
[25] I would thus not attempt to, as David Wiggins (1976, 138) writes, "evade responsibility by contriving [my] own fission."

be virtuous people too. If I did not repent for killing the innocent old lady, and Lefty and Righty each woke up in their respective beds creepily chuckling about how fun it was to kill the innocent old lady, then again my division would arguably involve an increase in overall fault. But in *Killer* I do repent; Lefty and Righty would thus not approve of killing the innocent old lady. To simplify matters, we should focus on cases of division in which no one is culpable for the division itself and in which no errors are made by the products of division, that is, we should focus on cases of division that involve no increase in fault.

Despite the fact that division does not per se involve any increase in fault, someone might find the Multiplication Argument for Division Multiplies Desert to be more plausible than the Fault Restriction. But before concluding that this is true, we should take note of a further implication of accepting the Multiplication Argument for Division Multiplies Desert: *indefinite* division would multiply desert indefinitely. Even if the total amount of deserved punishment could increase *somewhat* without any increase in fault, it might seem harder to accept that it could increase indefinitely without any increase in fault. Consider another case:

> ***Repeated Division.*** Technologically advanced hospitals have GrowthRay3000s, which can be used to stimulate pieces of brains to grow into whole brains exactly similar to those whole brains of which they were originally parts. GrowthRay3000s can bring about such desired effects within one three-thousandth of a second.[26]
>
> Now consider a different possible continuation of *Killer's Division.* With the help of some GrowthRay3000s, Lefty and Righty had the missing halves of their brains regrown from the halves that they each retained. Both then had whole brains and were physically and psychologically related to me in exactly the way that people are ordinarily related to their past selves. If I could be more than one person (I could not), I would be both. After Lefty and Righty woke up the next morning, they each decided to go for a walk. Astonishingly, Lefty was struck by a drunk driver. And doubly astonishingly, so was Righty! The accidents again destroyed their legs and torsos, and split their brains into halves. These four halves were taken to the hospital, and each was regrown into a whole brain and transplanted into a cloned body. Then there were four people: *Left-Lefty, Right-Lefty, Left-Righty,* and *Right-Righty.* Amazingly, all were in separate accidents exactly similar to mine, Lefty's, and Righty's. And so on.

[26] I am here inspired by Jacob Ross's (n.d.) use of various science fiction rays.



> After one division, there are two people. After two divisions, there are four. After three, there are eight, and so on. After just ten divisions, there are 1,024 people.

If we accept the Multiplication Argument, then assuming that I deserve X amount of punishment in *Killer,* we would claim that Lefty and Righty each deserve X amount of punishment in *Killer's Division,* making a total of 2X deserved punishment. But this argument can be reiterated. If Lefty splits into Left-Lefty and Right-Lefty, and they each deserve X punishment (as implied by Irrelevance of Others), and Righty splits into Left-Righty and Right-Righty, and they also each deserve X punishment (as implied by Irrelevance of Others), then the total becomes 4X deserved punishment. After ten divisions, there will be 1,024X deserved punishment, in total. Indeed, there is *no limit* on how much total deserved punishment there could be, all stemming from one wrong act. Reiterating the Multiplication Argument after each division implies the conclusion that indefinite division multiplies desert indefinitely.

Some people might find the Fault Restriction hard to deny, and they might find the implications of the Multiplication Argument in cases like *Repeated Division* hard to believe. To avoid contradiction, these people might have to deny either Desert or Irrelevance of Others. However, others might be willing to drop the Fault Restriction and to accept the implications of the Multiplication Argument in *Repeated Division.* For instance, they might defensibly argue that the Fault Restriction seems plausible in *non*division cases, but that it loses its intuitive appeal when applied to division cases (similar to the way in which Desert Requires Identity loses its intuitive appeal in division cases). They could thereby accept Desert and Irrelevance of Others without contradiction. However, Desert and Irrelevance of Others together imply Division Multiplies Desert, and Division Multiplies Desert might have implausible implications in the context of personal fusion.

5. Fusion

First, some more science fiction.

Suppose that, rather than a brain, my mind is realized in a futuristic liquid metal. My mind is distributed uniformly across the liquid metal, which can form various shapes. Usually, I take a form that is very hard to distinguish from an actual living human body. (Think of the T-1000, the bad guy from *Terminator 2.*) It is possible that I will divide. If I divide, the humanoid shape my liquid metal usually takes will form a puddle and divide like an amoeba into *n* qualitatively identical puddles. The resulting puddles will be the same size as the original one (perhaps the original puddle becomes *n* times larger immediately before dividing); each will be qualitatively identical to the original one, as well as spatiotemporally continuous with it. Then, the puddles will each morph into the particular humanoid shape the original one had.

It is also possible that I will *fuse* with other liquid metal persons like me. Consider another liquid metal person who is exactly like me. Suppose we each morph into a puddle and then each split into a left puddle and a right puddle, making a total of four qualitatively identical puddles. L1 and R1 are my puddles, L2 and R2 are his. Seconds later, puddles L1 and R2 are destroyed. But seconds after that, R1 and L2 come together, forming a single, unified puddle. I have now fused with another liquid metal person into a single *fusion product* who is qualitatively identical to each of us *fusion ingredients.*[27]

[27] Fusion isn't just for liquid metal puddle people. Take two qualitatively identical human persons, Shlefty and Shrighty. Owing to sufficient redundancies in their brains, if each had just one cerebral hemisphere, each would continue on exactly as each would if each retained both hemispheres. There are four cerebral hemispheres: L1, R1, L2, and R2. Suppose R1 and L2 are destroyed, and L1 and R2 are immediately connected. Shlefty and Shrighty have fused into a single person who is qualitatively identical to each of them.

Parfit (1984, 298–99), Unger (1990, chap. 6), and McMahan (2002, 83) each offer a brief but plausible discussion of fusion.

Cases of Dissociative Identity Disorder arguably provide real-life examples of fusion when the individual alters (separate persons realized in one and the same body) are integrated via therapy into a single person. Radden (1996) offers an intriguing discussion of fusion cases involving Dissociative Identity Disorder, including a discussion of the moral responsibility and punishment of the products of such fusion (though for an

It might be impossible to fuse certain persons together; attempting to fuse an old cynical scrooge and a young optimistic philanthropist might result in a cacophonous nightmare, an entity that is arguably not a person. Or maybe the resulting entity is a person, but one who happens to have a very chaotic psychology. But for the cases of fusion under discussion here, we should imagine that the persons who would fuse are physically and psychologically similar enough that they are "fusion compatible." Indeed, I will assume that the relevant fusion ingredients are as close to qualitatively identical as possible. For instance, we might imagine fusion stories like this one (told from the perspective of the distant future):

> ***Angela and Barbara.*** Angela and Barbara were liquid metal persons. They each came into existence in 2951 and were raised in separate but qualitatively identical controlled environments.[28] Though their lives from 2951 to 2999 were rich and complex, they remained qualitatively identical. On New Year's Day 3000, they fused in the liquid metal way described above. The resulting person is Carol.[29]

In fusion cases like this one, I believe that the following two claims are true:

> ***Prudential Comparability.*** Fusion ingredients (for example, Angela and Barbara) should be just as prudentially concerned about what happens to the product of their fusion (for example, Carol) as they should be

excellent criticism of her view on punishing fusion products, see D. Shoemaker 2009, 234–35). While these arguably realistic cases of fusion are well worth exploring, the more hypothetical cases I discuss in the main text are preferable, for present purposes, for two reasons: First, my hypothetical cases of fusion *are* cases of fusion if there can be any such cases at all. It is more debatable whether Dissociative Identity Disorder integration cases are truly cases of fusion because it is more debatable whether there are truly *two* preintegration persons and because it is more debatable exactly how the integration process works (for example, if one alter were simply eliminated, this would not be a case of fusion but at most a case in which one person dies and another person doesn't). Second, in my hypothetical cases of fusion, *Prudential Comparability* (see below) seems very hard to deny. It is more dubious whether Prudential Comparability holds in cases of Dissociative Identity Disorder integration.

[28] Perhaps these environments were controlled in the way Truman's was in *The Truman Show.*

[29] Recall the brief discussion of the metaphysics of division from section 1. There I explained why, at least according to certain Reductionist Views, it is plausible that I would not be identical to either of the products of my division. For similar reasons, it is plausible that neither Angela nor Barbara would be identical to Carol.

> prudentially concerned about what happens to their future selves, and in just the same way.

And:

> ***Desert Comparability.*** Punishing a fusion product is, from the point of view of Desert, tantamount to punishing each of the fusion ingredients in the same way and to the same extent. For example, if Angela deserved a punishment and Barbara deserved a punishment of the same size, then Desert would be just as satisfied if they were each given this punishment prefusion as it would be if Carol were given this punishment postfusion.

Why should we believe these claims?

Recall what I claimed about division: I should be just as prudentially concerned about what happens to Lefty as I should be prudentially concerned about what happens to my future self, whether or not I am identical to Lefty. (Same goes for me and Righty.) What matters for rational prudential concern is a psychological or physical relation, rather than identity. Whatever the particular nature of this psychological or physical relation is, it holds between me and Lefty, and between me and Righty. Similarly, there is a class of personal fusion cases where this relation holds between the fusion product and each of her fusion ingredients (but not between the fusion ingredients). For example, in *Angela and Barbara,* Carol is qualitatively identical to Angela, and the two are psychologically and spatiotemporally continuous. Indeed, the relations between Angela and Carol are exactly the same as the relations between Angela and future Angela, except that (i) Angela and Carol are nonidentical (on certain views), and (ii) Carol is the continuer of two people, whereas future Angela is the continuer of just one person. But neither (i) nor (ii) seem to *matter.*[30] That is, Angela's fusing with Barbara into Carol seems just as prudentially good for Angela as ordinary survival, and pun-

[30] One might note that, whereas future Angela is psychologically and physically continuous only with past Angela, Carol is psychologically and physically continuous with both Angela and Barbara. Could this difference plausibly ground the claim that (ii) does matter after all, and that, for example, Angela should have less prudential concern for Carol than she should have for future Angela?

I do not think so. Consider a pair of cases, again assuming in each case that, owing to sufficient redundancies in your brain, if you lost one cerebral hemisphere, you would continue existing exactly as you would if you retained both hemispheres.

In the first case, your left hemisphere is destroyed and then immediately replaced with an exact replica materialized from scratch. In this case, you survive, and everything that matters prudentially is preserved. In the second case, your left hemisphere is destroyed and then immediately replaced with an exact replica that belongs to a person ishing Carol seems just as prudentially bad for Angela as punishing future Angela.

Now, assuming that it is true that nonidentity holds between Angela, Barbara, and Carol, it is true that making Carol worse off to degree X might not, for example, make *Angela* worse off. However, we can say that making Carol worse off to degree X makes Angela *quasi*-worse off to degree X, or just "q-worse off" to degree X, in the following sense: if, while holding everything else constant, Barbara were taken out of this case, Angela *would* have been made worse off by degree X.

The difference between Angela's being made worse off to degree X and her being made q-worse off to degree X does not, from the point of view of her prudential concern, matter.[31] In the sorts of fusion cases under consideration here, it seems hard to deny Prudential Comparability. And Desert Comparability is true, I believe, in these cases in which Prudential Comparability is true. Consider the following argument:

(1) There is desert-based reason to make culpable wrongdoers A and B each worse off to degree X. (Desert)

(2) If there is desert-based reason to make A and B each worse off to degree X, then there is just as much desert-based reason to make A and B each q-worse off to degree X.

(3) There is just as much desert-based reason to make A and B each q-worse off to degree X (as there is to make A and B each worse off to degree X). (1 & 2)

(4) Making the fusion product of A and B worse off to degree X makes A and B each q-worse off to degree X.

(5) There is just as much desert-based reason to make the fusion product of A and B worse off to degree X (as there is to make A and B each worse off to degree X). (3 & 4)

qualitatively identical to you who has existed with this left hemisphere for exactly as long as you have but in a different part of the world.

It seems to me that what happens in the second case is prudentially no worse for you than what happens to you in the first case. The *history* of your left hemisphere replacement seems irrelevant. What seems prudentially relevant is what the left hemisphere replacement is going to do from now on.

[31] Unsurprisingly, the notion of being made q-worse off also applies to division cases. Consider a division case in which Lefty and Righty will live qualitatively identical lives and will each be made worse off to degree X at some point in the future. Insofar as I am not identical to either Lefty or to Righty, this might not make *me* worse off to degree X; but it would make me q-worse off to degree X. If, while holding everything else constant, Righty were taken out of this case, I *would* have been made worse off to degree X.

Assuming (1), we can avoid (5) only if we deny either (2) or (4). But (4) follows from the meaning of *q-worse off*: suppose the fusion product of A and B is made worse off to degree X; A would have been made worse off to degree X if B were taken out of the case, and B would have been made worse off to degree X if A were taken out of the case.

This leaves (2). The intuition undergirding (2) is that the difference between being made worse off and being made q-worse off should not matter to Desert since it is only a technicality involving the logic of identity that, in certain cases, forces us to say "q-worse off" rather than "worse off." It cannot be that there is desert-based reason to make someone worse off, but less or no desert-based reason to make that person q-worse off. (2) is intuitively plausible.

Thus, it seems that we cannot plausibly avoid (5). But (5) is simply another way of formulating Desert Comparability, the claim that Desert would be just as satisfied if each fusion ingredient were punished to degree X (or made worse off to degree X) as it would be if their fusion product were punished to degree X (or made worse off to degree X).[32] Punishing someone I prudentially should be concerned about in just the same way and to just the same extent that I prudentially should be concerned about my future self seems to be just as good, from the point of view of Desert, as punishing me. Of course, it is not the *same* as punishing me. Those who accept Desert Requires Identity will deny Desert Comparability. But, as I already argued, Desert Requires Identity seems false. Identity does not seem to be what matters for Desert.

For these reasons, in the sorts of fusion cases under consideration here, it seems hard to deny Desert Comparability. (I will now, for convenience, omit the qualification "in the sorts of fusion cases under consideration here.")

Next consider the "mirror image" of Division Multiplies Desert:

> ***Fusion Divides Desert.*** If *n* people who each deserve *m* years of punishment fuse, the fusion product deserves *m* years of punishment. The total prefusion punishment deserved is *m* times *n* years, and the total postfusion

[32] Note that (5) is formulated in terms of desert-based *reasons*, whereas Desert (and thus Desert Comparability) is formulated in terms of what *ought* to be done. But the claims about desert-based reasons in (5) imply the relevant oughts when considerations besides desert are held constant, or when these other considerations are not sufficiently weighty.

> punishment deserved is *m* times *n* years divided by *n*—or just *m* years. (Hence, fusion divides desert.)[33]

There is a simple and powerful argument for Fusion Divides Desert. According to Desert Comparability, giving the fusion product *m* years of punishment is as good, from the point of view of Desert, as giving each of these *n* fusion ingredients *m* years of punishment. That is, giving the fusion product *m* years of punishment is *desert comparable* to giving each of these *n* fusion ingredients *m* years of punishment. If so, giving the fusion product more than *m* years would seem to be desert comparable to overpunishing the fusion ingredients, and giving the fusion product fewer than *m* years would seem to be desert comparable to underpunishing the fusion ingredients.[34]

[33] An immediate worry one might have about Fusion Divides Desert is analogous to the worry about Division Multiplies Desert shared by those who believe in the Fault Restriction. Suppose there are one thousand murderers (who each have repented and are now virtuous) and that they each deserve twenty years of punishment. They form into puddles and fuse into one. According to Fusion Divides Desert, this one person deserves only twenty years. That is, the total amount of punishment deserved has decreased dramatically (divided by one thousand) just in virtue of an accident—not in virtue of any change in fault. Some might take this to be a good enough reason to reject Fusion Divides Desert. They might accept the

> *Fault Restriction**. There cannot be less deserved punishment if there is no decrease in fault.

However, just as one might defensibly claim that the Fault Restriction seems plausible in nondivision cases but not in division cases, one might defensibly claim that the Fault Restriction* seems plausible in nonfusion cases but not in fusion cases.

[34] Furthermore, notice that if we accept Division Multiplies Desert, we *might* be forced to also accept Fusion Divides Desert. Why? Consider

> *Killer's Division and Fusion.* Suppose I deserve twenty years of punishment for killing an innocent old lady. I divide into Lefty and Righty. Seconds later, Lefty and Righty fuse. Call the product of their fusion Feron. (I will here leave it open whether or not I am identical to Feron.) According to Division Multiplies Desert, Lefty and Righty each deserve twenty years of punishment. But it seems implausible that Feron would deserve anything other than twenty years of punishment.

Unless we can capture the claim that Feron deserves twenty years of punishment without appealing to Fusion Divides Desert, it seems implausible to accept Division Multiplies Desert without also accepting Fusion Divides Desert.

Moreover, a modified version of *Killer's Division and Fusion* provides further evidence against Desert Requires Identity. Suppose I have killed no one, and that Lefty and Righty each commit a murder during the few seconds they exist, before they fuse together into Feron. According to Desert Requires Identity, Feron cannot deserve punishment for what Lefty and Righty did, since he is identical to neither. But this seems implausible.

6. A Problem for Division Multiplies Desert

The problem, or puzzle, will not be apparent for several paragraphs. It takes some time to set it up. First, consider the fusion of an innocent person and a fully culpable murderer.

> ***Angela the Murderer.*** Remember *Angela and Barbara,* but now suppose that, on New Year's Eve 2999, Angela committed murder and Barbara did not. Instead, Barbara innocently observed a pretty sunset. Barbara deserves zero years of punishment, whereas Angela deserves twenty years. Then, on New Year's Day 3000, they fused into Carol.

What to do in tragic cases like this one, where an innocent person and a murderer fuse? On the one hand, we do not want the murderer to, well, get away with murder, and on the other hand, we do not want to do what is desert comparable to punishing the innocent person.

A reasonable response to *Angela the Murderer* is that it is, as it stands, underdescribed. Whether or not we should punish Carol, and how much, seems to depend on the details of how Angela and Barbara fused. After all, this new case is importantly different from *Angela and Barbara* since in the latter the fusion ingredients are qualitatively identical. But in *Angela the Murderer* the fusion ingredients are qualitatively different, owing to the fact that one but not the other committed murder—how, then, do these qualitatively different entities come together in fusion?

In particular, we might think it matters whether Carol has Angela's memory of committing murder, or instead has Barbara's memory of innocently observing a pretty sunset (that is, the memory of *not* committing murder).[35] We might also think it matters whether Carol *identifies with* Angela's act of murder, where this involves Carol embracing this act as her own, as an act that would intelligibly result from her central beliefs, desires, intentions, and personality traits.[36] In order to fill in such important details, I will add the following:

> ***Addendum to Angela the Murderer.*** Not long after Angela committed murder and Barbara observed a sunset, but before they fused, Barbara's mental life was altered, using a HypnoRay3000, to make it phenomenologically indistinguishable from Angela's mental life. Thus Barbara now has mem-

[35] Or, if memory presupposes personal identity, we can instead say that Carol has the *quasi*-memory of Angela's act of murder.

[36] See Schechtman 1996 and D. Shoemaker 2009, 220–28. The relation "person P identifies with action A" is not always one-to-one. In *Killer's Division,* for example, Lefty and Righty would both identify with, or *own,* my act of killing the innocent old lady.

> ories of and attitudes about Angela's act of murder that are phenomenologically just like Angela's; of course, these newly acquired memories of Barbara's are *false* memories. This alteration did not need to be very extensive, given that Angela and Barbara were qualitatively identical prior to their different behaviors on New Year's Eve. Angela deserves twenty years of punishment. Barbara deserves zero years of punishment. As before, they will fuse into Carol on New Year's Day.

Some people might object, claiming that Barbara deserves more than zero years of punishment, in virtue of what her mental life is now like (it is now phenomenologically just like Angela's). But I find this view hard to accept. The reason I find this view hard to accept is not that Barbara is not identical to Angela, the wrongdoer; recall that there are powerful reasons for rejecting Desert Requires Identity. Rather, the reason I find it hard to accept is that Barbara's current mental life, though phenomenologically just like Angela's, does not have the appropriate sort of *cause.* Barbara's current mental life is phenomenologically just like a murderer's because of a HypnoRay3000 alteration. That is not the appropriate sort of cause to ground desert. Angela's current mental life is phenomenologically just like a murderer's because she committed murder. That is the appropriate sort of cause to ground desert. In *Killer's Division*, Lefty's mental life is phenomenologically just like a murderer's because he is the continuer of someone who committed murder. That too is the appropriate sort of cause. These claims are plausible.

The view that Barbara deserves any more than zero years of punishment thus seems implausible. At the very least we should agree that, due to the fact that *a* particularly important sort of cause is missing in the case of Barbara that is present in the case of Angela, Barbara deserves significantly *less* than twenty years of punishment. That is, we should at least agree that *Angela the Murderer*, with the Addendum, is a fusion case in which the fusion ingredients deserve significantly different punishments.

It is hard to say what to do in tragic cases like *Angela the Murderer.* I suspect that we can, however, make some progress by stepping back and thinking about overpunishing and underpunishing in cases that involve neither fusion nor division. Consider

> ***Differentially Deserving.*** There are *n* people who each deserve twenty years of punishment and one person who deserves only fifteen years of punishment. For some reason, we cannot give everyone exactly what they deserve. We have to either punish no one at all, or else give everyone the same amount of punishment, X, where X is between twenty years of

> punishment and fifteen years of punishment. We are again setting aside benefits that might come from punishing.

First, suppose that $n = 1$. (Later I will return to cases where n is greater than 1.)

There are different views we could take about what X should be in cases like *Differentially Deserving.* We could accept

> ***Average Punishment.*** Since the reasons for giving each person a fitting punishment are equally strong, in cases where we must give each the same punishment, we should give each the average of what they deserve.

Average Punishment would imply that, in *Differentially Deserving,* assuming $n = 1$, X should be 17.5. Setting X higher than 17.5 would reflect the belief that there are stronger reasons to give the person who deserves twenty years a fitting punishment (that is, the punishment he or she deserves), and setting X lower than 17.5 would reflect the belief that there are stronger reasons to give the person who deserves fifteen years a fitting punishment. But since, on this view, there are equally strong reasons to give each a fitting punishment, X should be 17.5. Alternatively, we could accept

> ***Weighted Average Punishment.*** Since the reasons for giving those who deserve less (or no) punishment a fitting punishment are stronger than the reasons for giving those who deserve more punishment a fitting punishment, in cases where we must give each the same punishment, we should give each a *weighted* average of what they deserve. In determining the average, greater weight is placed on what those who deserve less punishment deserve.

Weighted Average Punishment would imply that, in *Differentially Deserving,* assuming $n = 1$, X should be *less* than 17.5. How much less than 17.5? That depends on how much extra weight is placed on giving the less deserving of punishment (or the innocent) what they deserve. For example, some might accept something like Blackstone's Formula that "it is better that ten guilty persons escape than that one innocent suffer" (Blackstone 1915 [1765], 523). It is possible to claim that we should give *absolute* weight to giving the less deserving of punishment (or the innocent) what they deserve. That is, we could accept

> ***Least Punishment.*** Since the reasons for giving those who deserve less (or no) punishment a fitting punishment are *absolutely* stronger than the reasons for giving those who deserve more punishment a fitting punish-

> ment, in cases where we must give each the same punishment, we should give each what the person deserving of the least punishment deserves.

Least Punishment would imply that, in *Differentially Deserving*, assuming $n = 1$, X should be 15. This view seems implausibly extreme. Moreover, independently of how intuitively plausible or implausible Least Punishment is, if defenders of Desert accepted it, their view would imply that it would virtually always be too risky to justifiably engage in punishment. This is because, whenever any punishment is carried out, there is virtually always some nonzero *risk* of giving someone more punishment than they deserve. But I assume that defenders of Desert do think—or at least would like it to be the case—that their view implies that in many cases it is *not* too risky to engage in punishment. For this reason, and because Least Punishment is implausibly extreme, I assume that defenders of Desert would wisely reject it.

In cases like *Differentially Deserving*, where $n = 1$, it seems defenders of Desert must choose between Average Punishment and Weighted Average Punishment.

So far, we have only been considering versions of *Differentially Deserving* in which $n = 1$. We can now ask what should happen as n increases. Both Average Punishment and Weighted Average Punishment imply that as n increases, X should increase. We might accept Average Punishment or Weighted Average Punishment in cases where $n = 1$, but deny these views in some cases where n is greater than 1. I will not explore all the possible views here. A rather modest view, which it seems defenders of Desert must accept, is:

> ***Numbers Matter.*** There is some number n such that X should be greater than what X should be if instead n were 1.

If defenders of Desert denied this, they would be claiming that, in cases where,

> (1) we must give *an arbitrarily large number* of people who each deserve twenty years of punishment and one person who deserves fifteen years of punishment the same amount of punishment,

we should assign the same punishment, that is, the same size for X, as in cases where,

> (2) we must give *one* person who deserves twenty years of punishment and one person who deserves fifteen years of punishment the same amount of punishment.

But it seems hard to believe that if we accept Desert, X should be the same size in both (1)-cases and in (2)-cases. This would seem not to give due weight, or any weight, to the reasons for giving the extra people in (1)-cases who deserve more punishment their fitting punishments (twenty years). Since these are reasons that I believe defenders of Desert must claim exist, I believe that they cannot plausibly deny Numbers Matter. This concludes my discussion of overpunishing and underpunishing in cases that do not involve fusion or division.

Recall that, according to Desert Comparability, punishing a fusion product is, from the point of view of Desert, tantamount to punishing each of the fusion ingredients. With this in mind, consider two analogues of two claims mentioned above:

> ***Average Punishment*** $_{\text{FUSION}}$. The product of fusion deserves the amount of punishment that is the average of the amounts of punishment deserved by each of the fusion ingredients.
>
> ***Weighted Average Punishment*** $_{\text{FUSION}}$. The product of fusion deserves the amount of punishment that is the weighted average of the amounts of punishment deserved by each of the fusion ingredients. In determining the average, greater weight is placed on what those fusion ingredients who deserve less punishment deserve.

Next recall that Average Punishment, Weighted Average Punishment, and Numbers Matter apply to cases where we must either give each person the same punishment or give each no punishment at all. In fusion cases, we must do what is *desert comparable* to giving each of the fusion ingredients the same punishment or else giving each no punishment at all.[37] Desert Comparability thus implies that fusion cases are, in certain relevant ways, like cases where we must give each person the same punishment. Indeed, it seems that:

[37] As already seen in the brief discussion of Fusion Divides Desert, this is why (again, at least in the relevant sorts of fusion cases) it would be implausible to *add up* the punishments of the fusion ingredients, in determining how much the fusion product ought to be punished. Below I discuss *Ordinary Fusion*, a case in which Angela deserves twenty years of punishment, Barbara deserves fifteen years of punishment, and they fuse into Carol (after the relevant HypnoRay3000 alteration of Barbara's mental life). It would be implausible to claim that Carol ought to receive thirty-five years of punishment. This is because giving Carol thirty-five years of punishment would be desert comparable to *both* giving Angela thirty-five years of punishment *and* giving Barbara thirty-five years of punishment, which would be significantly overpunishing each of them.

- Average Punishment and Desert Comparability together imply Average Punishment $_{\text{FUSION}}$.
- Weighted Average Punishment and Desert Comparability together imply Weighted Average Punishment $_{\text{FUSION}}$.
- Numbers Matter and Desert Comparability together imply *Numbers Matter* $_{\text{FUSION}}$ (illustrated below).

Suppose that *one million* murderers, who each deserve twenty years of punishment, fused with one murderer who deserves fifteen years of punishment. Assuming Desert, it seems implausible that the product of this fusion deserves only fifteen years of punishment. Even 17.5 years would seem too lenient. Further, let

> P_1 = the product of the fusion of one murderer who deserves twenty years of punishment and one murderer who deserves fifteen years of punishment,

and let

> P_2 = the product of the fusion of *n* murderers who each deserve twenty years of punishment and one murderer who deserves fifteen years of punishment.

It seems quite hard to deny the following modest claim about P_1 and P_2:

> ***Numbers Matter*** $_{\text{FUSION}}$. There is some finite number *n* such that P_2 deserves more punishment than P_1. (I will assume that *n* is sufficiently large if it is at least one million.)

I believe we are now finally in a position to see that Division Multiplies Desert faces a potentially serious problem.

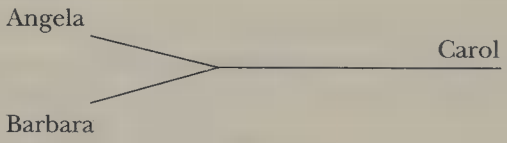
In *Angela the Murderer*, Angela deserves twenty years of punishment. But now suppose that Barbara was also a murderer and that she deserves fifteen years of punishment. Barbara's act of murder is deserving of less punishment than Angela's, we can suppose, because whereas Angela's murder was premeditated, Barbara's was more spur-of-the-moment. Also suppose, as we did in the Addendum, that Barbara's mental life was altered using a HypnoRay3000 to make it phenomenologically indistinguishable from Angela's mental life. Thus Barbara now has memories of, and attitudes about, Angela's act of premeditated murder that are phenomenologically just like Angela's; of course, these newly acquired memories of Barbara's are *false* memories—she now falsely remembers planning out the murder and embraces this premeditation as her own.

Even though Barbara's mental life is now phenomenologically just like Angela's, there is again a relevant causal difference. Barbara's mental life is *partly* appropriately caused, as she has the memory that she committed murder because she committed murder. But her mental life is also *partly not* appropriately caused, as she has the memory that she committed murder *with premeditation* because of a HypnoRay3000 alteration. Angela's current mental life, by contrast, is *wholly* appropriately caused. She has the memory that she committed murder with premeditation because she committed murder with premeditation.

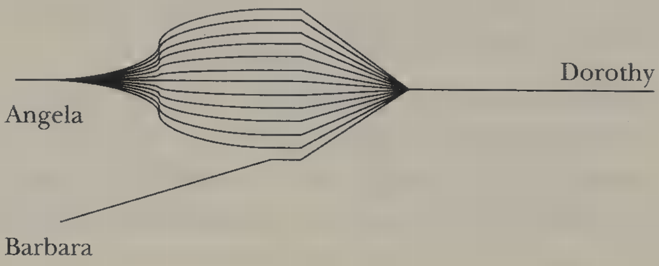
For these reasons, it is plausible that, whereas Angela deserves twenty years of punishment, Barbara deserves only fifteen years of punishment. At the very least, we should agree that Barbara deserves significantly *less* than twenty years of punishment.

Now compare *Ordinary Fusion* with *First Angela Divides*:

Ordinary Fusion



First Angela Divides

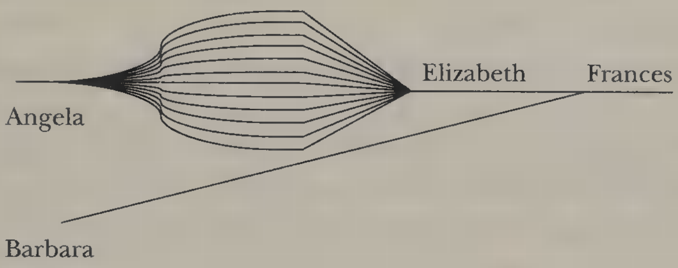


In *Ordinary Fusion*, Angela and Barbara fuse. The resulting person is Carol. In *First Angela Divides*, Angela divides into one million persons. (Diagram is not to scale.) Seconds later, these one million persons, and Barbara, fuse. The resulting person is Dorothy.

According to Division Multiplies Desert, each of the one million products of Angela's division deserves twenty years of punishment. But then, according to Numbers Matter $_{\text{FUSION}}$, Dorothy deserves more punishment than Carol. But this seems implausible.

If Angela's division resulted in an increase in fault, by creating many more people who are vicious or who fail to repent, then it would not be as implausible to claim that Dorothy deserves more punishment than Carol. But here, as before, we are restricting our focus to cases of faultless division. Given this, it seems implausible that Dorothy deserves more punishment than Carol. Division per se cannot make this kind of difference. The following case reveals further evidence that Carol and Dorothy deserve the same amount of punishment:

First Angela Divides and Fuses



In *First Angela Divides and Fuses,* Angela divides into one million persons, which, seconds later, fuse. Call the product of this fusion Elizabeth. Then, seconds after that, Elizabeth fuses with Barbara. Call the product of this final fusion Frances.

Since Elizabeth should deserve the same amount of punishment as Angela,[38] Carol and Frances should deserve the same amount of punishment. (Because Carol is the product of the fusion of Barbara and Angela, and Frances is the product of the fusion of Barbara and someone—Elizabeth—who deserves just as much punishment as Angela.) Moreover, it seems plausible that Frances and Dorothy should deserve the same amount of punishment. After all, it does not seem to matter, morally, whether Barbara fuses with the products of Angela's division *while* they are fusing or *after* they have fused into Elizabeth. This, in turn, provides further support for the claim that Carol and Dorothy deserve the same amount of punishment. (Because Carol and Frances deserve the same amount of punishment.)

To avoid the implausible implication that Dorothy deserves more punishment than Carol, we must deny either Numbers Matter $_{\text{FUSION}}$ or Division Multiplies Desert. Numbers Matter $_{\text{FUSION}}$, I claimed, is hard to

[38] Recall *Killer's Division and Fusion,* presented in note 34.

deny if we accept Desert. This is because Numbers Matter is hard to deny if we accept Desert, and Numbers Matter and Desert Comparability together imply Numbers Matter $_{\text{FUSION}}$. We can now state the

> ***Fusion Problem.*** We must deny either
> (i) Dorothy deserves no more punishment than Carol; or,
> (ii) Numbers Matter $_{\text{FUSION}}$; or,
> (iii) Division Multiplies Desert

Here are two possible solutions to the problem.

The first possible solution is that, while Numbers Matter $_{\text{FUSION}}$ is plausible in less exotic cases of fusion in which the one million persons who each deserve twenty years of punishment did not result from division, it is implausible in cases like *First Angela Divides.* However, this solution seems implausible *if* it is true—as is implied by Division Multiplies Desert—that each of the products of Angela's division *really* does deserve twenty years of punishment and that there is as much reason to punish each of them as there is to punish Angela. To avoid letting these many division products off too easy, we have to punish Dorothy more. Analogously, it would be implausible to deny Numbers Matter in versions of *Differentially Deserving* in which the large number *n* of persons who deserve more punishment are each products of the division of a single person who deserves more punishment.

The second possible solution is that, while Division Multiplies Desert is plausible in less exotic cases in which division is not followed seconds later by fusion, it is implausible in cases like *First Angela Divides* (in which division is followed seconds later by fusion). This solution implies that how much punishment is deserved by the products of Angela's division depends on whether they will later fuse. But suppose we had the power to cram the equivalent of twenty years of punishment into a short span of time, such that we could punish the products of Angela's division in *First Angela Divides* before they fuse. And suppose it is important that we carry out punishment within this critical window (perhaps because we will very soon lose the ability to punish at all). According to the second solution, we ought to punish the products of Angela's division if we knew they would not later fuse, but we ought not to punish the products of Angela's division in *First Angela Divides* (or we ought to punish them, but to a considerably lesser extent). This is quite implausible.

It is very puzzling that Dorothy should deserve any more punishment than Carol. But Division Multiplies Desert implies that, in *First Angela Divides,* each of Angela's division products deserves just as much punishment as she does. This generates a reason to do what is desert comparable to punishing each of these division products for twenty years. But this reason is absent in *Ordinary Fusion.* So, puzzlingly, we seem to have a reason to punish Dorothy more than Carol. There does not appear to be a non-implausible solution to the Fusion Problem: to avoid a contradiction, we must deny (i), (ii), or (iii). Denying (i) seems implausible, and it seems hard to deny (ii), at least if we accept Desert. We could instead deny (iii), Division Multiplies Desert. But as explained earlier, denying Division Multiplies Desert would require either denying Desert or denying Irrelevance of Others. But, as was pointed out in section 3, it seems implausible to deny Irrelevance of Others.

7. Rethinking Desert?

We began with *Killer.* Many believe that, at least in that case, I ought to be punished for what I did. Then we considered *Killer's Division* and the Multiplication Argument. The Multiplication Argument supports the following conclusion: three beliefs about desert are inconsistent. These beliefs are Desert, Irrelevance of Others, and Irrelevance of Division. If we do not imagine the right kinds of cases, or do not reflect carefully about them, we will not notice that these three beliefs are inconsistent. But, as my argument showed, they are inconsistent. We might have thought that we can defensibly deny Irrelevance of Division, while maintaining Desert and Irrelevance of Others. However, Desert and Irrelevance of Others entail Division Multiplies Desert. If we find the Fault Restriction plausible, we will resist Division Multiplies Desert, and we will perhaps find it repugnant that indefinite division would multiply desert indefinitely. But, I claimed, it seems that the Fault Restriction can be defensibly denied. However, if we accept Division Multiplies Desert, we might face serious problems in the context of personal fusion. In particular, if we accept Division Multiplies Desert, then we will have to implausibly deny either (i) or (ii) in the Fusion Problem.

I have done more, in this essay, than raise difficult questions about how to work out the *implications* of Desert in division and fusion cases. I might have done no more than this if, for example, the only puzzle raised here was about how or whether we ought to punish the fusion products of differentially deserving fusion ingredients. But I believe that, in addition to raising difficult questions about how to work out the implications of Desert, I have given us some reason to rethink Desert itself.

I have shown that, to maintain Desert, we have to deny either Irrelevance of Others, or (i) or (ii) in the Fusion Problem. Giving due weight to the independent plausibility of Irrelevance of Others and (i) and (ii) in the Fusion Problem yields at least *some* reason to rethink Desert. A *possible* conclusion to draw is that the division and fusion cases I have discussed here reveal the fact that we simply ought to deny Desert. It is not clear that this is the right conclusion to draw, but it does seem to be a mistake to claim that it is easily avoided, or avoided at no cost.

This conclusion could not plausibly be avoided, for example, by baldly stipulating that Desert *does not apply* to division and fusion cases. Such a restriction would itself violate Irrelevance of Others. *Only Lefty Survives* is a nondivision case. So Desert would imply that Lefty deserves punishment. But if Righty survived too, we would have a division case, and so Desert would no longer apply, and would thus not imply that either deserves punishment. That is implausible.

We could perhaps rethink and even deny Desert without giving up on other desert (lowercase "d") views, according to which people deserve to be punished independently of, or on top of, the benefits that may result from punishment (for example, through deterrence).[39] Recall that Desert specifically claims that, other things being equal, when people culpably do very wrong or bad acts, they deserve punishment in the sense that they ought to be made worse off simply in virtue of the fact that they culpably did wrong, even if they have repented, are now virtuous, and punishing them would benefit no one. Focusing on a specific view like Desert, rather than desert views in general, made it easier to investigate how some desert considerations might play out in division and fusion cases. But the truth is that the conceptual landscape of desert is complex.[40] And the intersection of desert and division and fusion cases may thus be exceedingly rich. This essay reveals, at most, the tip of the iceberg.

Finally, personal division and fusion cases might not be puzzling only for Desert and its kin, but also for a wide variety of normative views that are relevantly bound up with personal identity; for example, views about the debts that persons owe to others, views about the just distribution of benefits and harms across separate persons, and views about

[39] I am thus not counting utilitarian accounts of punishment as bona fide desert views. See, for example, Smart 1961 and Arneson 2003.

[40] Shelly Kagan's work is a testament to this fact (see Kagan 2012b).

special partial concern for oneself or one's intimates.[41] While I am somewhat skeptical that analogues of the division and fusion puzzles I have discussed here will *threaten* all such views in the way I think they might threaten Desert,[42] these remain early days. It is at least *possible* that carefully designed examples and arguments will reveal many views that share certain structural features with Desert to be implausible, or less plausible than they seemed prior to considering personal division and fusion.[43]

References

Arneson, Richard. 2003. "The Smart Theory of Moral Responsibility and Desert." In *Desert and Justice*, ed. Serena Olsaretti, 233–58. Oxford: Oxford University Press.

Blackstone, William. 1915 [1765]. *Blackstone's Commentaries.* Abridged by William Sprague. 9th ed. Chicago: Callaghan and Company.

Campbell, Tim. n.d. "Should Utilitarians Care about People?" Unpublished manuscript.

Chisholm, Roderick. 1976. *Person and Object: A Metaphysical Study.* La Salle, IL: Open Court.

[41] Jacob Ross (n.d.) and Tim Campbell (n.d.) have written excellent papers on the implications of fission and fusion cases for ethics. Ross's paper focuses on special concern, and Campbell's focuses on aggregation and the bearers of value. Their distinct and interesting puzzles are importantly structurally different from mine.

[42] Perhaps, for instance, a "Divided Debt" view, according to which each of *N* division products would owe 1/*N* of the debt that their debtor ancestor owes to some debtee, would not be implausible in the way that Divided Desert (from section 3) is implausible. After all, if matters of luck, like whether a debtee spontaneously decides to forgive a debtor's debt, or whether a third party spontaneously decides to incur a debtor's debt, can plausibly reduce a debtor's debt (and winning the lottery can make it very easy to repay debt), then perhaps "division-luck," including how many "division siblings" a product of a debtor's division happens to have, can likewise plausibly reduce the debt owed by each of the debtor's division products. By contrast, luck cannot plausibly affect how much punishment a person *deserves* to suffer. This difference may explain why Irrelevance of Others cannot plausibly be denied, while its debt analogue might reasonably be denied.

[43] This essay has a companion piece (Pummer, n.d.) titled "Fission, Fusion, and the Ethics of Distribution." In it, I argue that fission and fusion cases put serious pressure on two widely held views in distributive justice, or the ethics of distribution: that it matters how well off people are relative to others, and that it is easier to justify balancing benefits and harms that occur within the life of a single person than it is to justify balancing benefits and harms that occur within the lives of separate persons. I tentatively conclude that if we cannot resolve certain puzzles raised by fission and fusion cases, we might have reason to rethink the normative significance of the separateness of persons, and (more radically) we might have reason to claim that, while it matters whether there *are* more or fewer benefits and harms, it does not ultimately matter *who* receives them.

DeGrazia, David. 2005. *Human Identity and Bioethics.* Cambridge: Cambridge University Press.

Johnston, Mark. 1989. "Fission and the Facts." *Philosophical Perspectives* 3: 369–97.

Kagan, Shelly. 2012a. *Death.* New Haven: Yale University Press.

———. 2012b. *The Geometry of Desert.* New York: Oxford University Press.

Lewis, David. 1976. "Survival and Identity." In *The Identities of Persons,* ed. Amélie Rorty, 17–40. Berkeley: University of California Press.

Locke, John. 1975 [1694]. *An Essay concerning Human Understanding,* ed. Peter Nidditch. Oxford: Oxford University Press.

McMahan, Jeff. 2002. *The Ethics of Killing.* New York: Oxford University Press.

Merricks, Trenton. 1998. "There Are No Criteria of Identity Over Time." *Noûs* 32: 106–24.

Nozick, Robert. 1981. *Philosophical Explanations.* Cambridge, MA: Harvard University Press.

Olson, Eric. 1997. *The Human Animal: Personal Identity without Psychology.* Oxford: Oxford University Press.

Parfit, Derek. 1971. "Personal Identity." *Philosophical Review* 80: 3–27.

———. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

———. 1986. "Comments." *Ethics* 96: 832–72.

Pereboom, Derk. 2001. *Living Without Free Will.* Cambridge: Cambridge University Press.

Perry, John. 1972. "Can the Self Divide?" *Journal of Philosophy* 69: 463–88.

Pummer, Theron. n.d. "Fission, Fusion, and the Ethics of Distribution." Unpublished manuscript.

Radden, Jennifer. 1996. *Divided Minds and Successive Selves.* Cambridge, MA: MIT Press.

Ross, Jacob. n.d. "Can Revisionary Metaphysics Save Commonsense Ethics." Unpublished manuscript.

Schechtman, Marya. 1996. *The Constitution of Selves.* Ithaca, NY: Cornell University Press.

Shoemaker, David. 2009. *Personal Identity and Ethics: A Brief Introduction.* Peterborough, ON: Broadview Press.

Shoemaker, Sydney. 1970. "Persons and Their Pasts." *American Philosophical Quarterly* 7: 269–85.

———. 1984. "Personal Identity: A Materialist's Account." In *Personal Identity,* ed. Sydney Shoemaker and Richard Swinburne, 67–132. Oxford: Blackwell.

Smart, J. J. C. 1961. "Free Will, Praise, and Blame." *Mind* 70: 291–306.

Swinburne, Richard. 1984. "Personal Identity: The Dualist Theory." In *Personal Identity,* ed. Sydney Shoemaker and Richard Swinburne, 1–66. Oxford: Blackwell.

Thomson, Judith. 1997. "People and Their Bodies." In *Reading Parfit,* ed. Jonathan Dancy, 202–29. Oxford: Blackwell.

Unger, Peter. 1990. *Identity, Consciousness, and Value.* New York: Oxford University Press.

Velleman, David. 2008. “Persons in Prospect.” *Philosophy and Public Affairs* 36: 221–88.

Wiggins, David. 1976. “Locke, Butler and the Stream of Consciousness.” *Philosophy* 51: 131–58.

Agreement Matters: Critical Notice of Derek Parfit, *On What Matters*

Stephen Darwall
Yale University

When Derek Parfit's *Reasons and Persons* (*RP*) appeared in 1984, it posed stiff challenges to moral philosophical orthodoxy across the board. Where the 1970s had been dominated by Kantian deontological moral and political theory, owing mainly to Rawls's influence, Parfit vigorously defended teleology in its most radical form: act consequentialism grounded in an "ultimate moral aim: that outcomes be as good as possible."[1] Rawls had held that satisfying a "publicity condition" was a "constraint of the concept of right" (Rawls 1971, 115). And Bernard Williams had argued that since the public acceptance of act consequentialism in place of moral common sense would have bad consequences, the theory must "usher itself from the scene" (Smart and Williams 1973, 134). Against such objections, Parfit argued that even if act consequentialism would have to remain an "esoteric" theory, as Sidgwick put it, that would not constitute "a ground for doubt" (Parfit 1984, 41). A moral theory might be true even if it is "indirectly self-defeating" in this way. Parfit thus challenged the widely held Rawlsian view that any plausible moral theory must be apt for inclusion in "public reason."

I am indebted to participants in a seminar on *On What Matters* at Yale in Fall 2012 and, especially, to Shelly Kagan, with whom I taught the course.

[1] Parfit defended 'C' ("There is one ultimate moral aim: that outcomes be as good as possible"), which when "applied to acts" yields "What each of us ought to do is whatever would make the outcome best" (Parfit 1984, 24).

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366495

A second Parfitian challenge concerned personal identity. Rawls had argued that the dignity of persons entailed the moral significance of "the plurality and distinctness of individuals" (Rawls 1971, 29). Whereas utilitarianism holds that a cost to A can be offset by the same benefit to B that would offset it for A, Kantian theories like Rawls's take the "separateness of persons" seriously. According to them, as well as, indeed, to moral common sense, it matters intrinsically that A and B are different people. *RP* struck at the metaphysical underpinnings of this idea, arguing that personal identity consists metaphysically in relations of physical and psychological continuity that hold to varying degrees rather than as a "deep further fact" (Parfit 1984, 280). If this is so, Parfit argued, utilitarians can plausibly argue that deontological distributive principles grounded in the "separateness of persons" should be given less weight (Parfit 1984, 337, 339–46).

Yet another challenge to deontological "person-affecting" principles was posed by Parfit's "non-identity" problem (Parfit 1984, 351–80). Suppose that we want to condemn unfettered use of fossil fuels on the grounds that this harms future generations. We may not be able to claim that the people who will actually exist will have been harmed. Changing patterns of energy consumption also change times of conception and consequently who ends up existing. Assuming that their lives are worth living, therefore, it seems that future individuals will not in fact have been harmed on balance by our profligate consumption. From this perspective, it can seem as if the only possible moral objection to current profligacy is the consequentialist complaint that it will make the world worse *impersonally*, not that it harms future individuals.

These are perhaps enough examples to illustrate two features of *RP* that are difficult to ignore while reading Parfit's recent massive work, *On What Matters* (*OWM*). First, in its method and style, *RP* had a bracing freshness that employed philosophical, often metaphysical, arguments to challenge apparently agreed moral views. It was a deeply antiestablishment work.[2] And second, the moral theory *RP* defended could hardly have been more anti-Kantian. Whereas Kantians like Rawls had held that "the right is prior to the good," Parfit defended a teleological theory according to which morality has "an ultimate moral aim." And he applied

[2] There are times when one almost has the sense that *RP* bears a relation to *OWM* that is something like that borne by Parfit's radical young Russian nobleman to his later more conservative self (Parfit 1984, 327).

this aim directly to acts: "what each of us ought to do is whatever would make the outcome best" (Parfit 1984, 24).

Read against this background, *OWM* is striking in a number of respects. Two of the most prominent are the book's sympathetic engagement with Kantian moral philosophy and defense of a rule- in place of an act-consequentialist theory of right, on the one hand, and its self-conscious search for points of convergence and agreement in moral theory, on the other. Parfit's central project in volume 1 is to argue that three major moral philosophical traditions: Consequentialism, Kantianism, and Scanlonian Contractualism, when given their philosophically best versions, converge on what he calls the

> Triple Theory: An Act is wrong if and only if, or *just when*, such acts are disallowed by some principle that is
>
> (1) one of the principles whose being universal laws would make things go best [Rule Consequentialism],
>
> (2) one of the only principles whose being universal laws everyone could rationally will [Kantian Contractualism],
>
> (3) a principle that no one could reasonably reject [Scanlonian Contractualism] (1:413).

Parfit's "Convergence Argument" attempts to prove that principles that would satisfy any one of these three formulae would also satisfy the other two. Proponents of these three major traditions have been, Parfit says, "climbing the same mountain on different sides" (1:xxiii).

Already we see important differences from *RP.* First, what Parfit evidently means by a principle's "being universal law" in clauses (1) and (2) of the Triple Theory is its *being universally accepted* (in common) as a principle of moral right and wrong (see, for example, 1:355). And a parallel point holds for (3). What is in question is the rational choice-worthiness or reasonable rejectability of people having certain beliefs in common about right and wrong.[3] Thus whereas Parfit insisted in *RP* that whether a moral theory should be accepted as common sense is irrelevant to the theory's truth, *OWM* apparently takes a different view.

What has led Parfit to this different view? The Triple Theory's formulae all, in effect, respect a Rawlsian "publicity condition" for which,

[3] Parfit presents the "Kantian Contractualist Formula" ("Everyone ought to follow the principles whose universal acceptance everyone could rationally will") as a lightly revised version of what he calls "Kant's Moral Belief Formula" ("It is wrong to act in some way unless *everyone* could rationally will it to be true that everyone believes such acts to be morally permitted") (1:20).

as Parfit had himself shown in *RP*, teleological act consequentialism has no use. Deontological theories like Rawls's and Kant's hold that the deontic concept of right has its own distinctive conceptual constraints that are independent of, and, Rawls says, "prior to," those of the good.[4] I shall argue that in *OWM*, Parfit effectively agrees with the Rawls/Kant tradition about this fundamental conceptual matter.

In *RP*, Parfit did not distinguish between what a person morally should do in the sense of what there is most moral reason to do and what it is morally obligatory or right in the sense of what it would be morally *wrong* not to do.[5] In *OWM*, however, Parfit distinguishes what he now calls the "deontic" concepts of moral right and wrong.[6] Chapter 7, "Moral Concepts," discusses the "*moral* senses of 'wrong', and the concepts that these senses express" (1:150). And though Parfit never directly affirms there the Rawlsian thesis that publicity is a constraint of the (deontic) concept of right (though not of the good), I shall suggest that he effectively concedes the point and that this helps explain Parfit's move from defending act consequentialism to defending rule consequentialism.

Why does the desire to be in agreement pervade *OWM*, as it did not *RP*? Partly Parfit is moved by the general epistemological problem of disagreement between epistemic peers. Parfit tells us that although he had initially intended to think further about the questions he investigated in *RP*, he "became increasingly concerned about certain differences between [his] views and the views of several other people" (2:427). He then quotes Sidgwick:

> If I find any of my intuitions in direct conflict with an intuition of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two intuitions necessarily reduces me . . . to a state of neutrality. (2:427)

OWM is consumed with these differences and disagreements, particularly with the views of philosophers Parfit mostly highly respects, like Bernard Williams. Volume 1 is concerned primarily with normative dis-

[4] I take Rawls's claim not to be that a theory of the right is somehow prior to a theory of good, but that when it comes to theorizing about moral right (and, of course, justice), one must begin with whatever conceptual constraints these latter concepts involve, such as the "publicity condition." These then constrain the way in which propositions and theories of the good figure in.

[5] See, for example, Parfit 1984, 25, on "objective rightness."

[6] The term did not appear in *RP*.

agreements, especially about moral right and wrong. Here Parfit argues that the disagreements are largely superficial; when given their philosophically best shape, contending theories of the right converge on the Triple Theory.

Volume 2, by contrast, focuses mostly on metaethics. Here, nothing like a Triple Metaethical Theory is in the offing. Parfit vigorously defends his metaethical position, nonnaturalist cognitivism, against noncognitivism, antirealist error theories, and all forms of ethical naturalism, not looking for points of convergence, but sometimes arguing that no real disagreement exists since some positions, like Williams's, do not employ genuine normative concepts.

Strikingly, Parfit's own views seem less destabilized by metaethical disagreement.[7] Disagreement with epistemic peers seems, however, to be no less a reason to doubt metaethical views than moral convictions. But perhaps we should be bothered less by metaethical disagreements than by comparable disagreements about moral right and wrong. Although disagreement with epistemic peers may properly lead us to wonder who has grasped the truth of some matter, it does not put metaethical truth itself in question. Either there are irreducible nonnatural normative facts or there are not. Sufficiently divergent moral opinion in optimal epistemic conditions, however, can lead us to wonder whether there are moral facts at all. The reason, I shall suggest, has to do with constraints on the concept of right that underlie Rawls's publicity principle.

On What Matters is clearly an important book. Whether it will have the impact and influence of *Reasons and Persons* remains, of course, to be determined. Many who were inspired by *RP* either to pursue consequentialist theories with renewed vigor or to work on the moral implications of the metaphysics of personal identity and the nonidentity problem, and the like, may find Parfit's strategy and conclusions in *OWM* somewhat disappointing. And likely some who were attracted by Parfit's reductionist metaphysics of personal identity will find themselves unsettled and perhaps confused by Parfit's nonreductionist impulses when it comes to the normative. There can be no doubt, however, that *OWM* represents the results of a fascinating intellectual journey by one of the greatest moral philosophical minds of recent times. Anyone seriously interested in systematic normative ethics or metaethics will have to engage with it. I predict that *OWM* will have significant impact, but that this may come slowly.

[7] Samuel Scheffler notes this difference in his introduction at 1:xxvi.

With 1,070 pages of main text, 157 pages of appendices, and 64 pages of substantial endnotes, there is quite a bit to work through. And it is not easy going. Parfit's prose, if complex, is famously lean and exact. So there is precious little fat. Volume 2 begins with commentaries by Susan Wolf, Allen Wood, Barbara Herman, and T. M. Scanlon, and these chapters by contrast sometimes seem almost to fly by.

Volume 1 begins with an elegantly written introduction by Samuel Scheffler. Scheffler lays out the argument of parts 1 through 5 (part 6 concerns the metaethics of normativity) and flags likely points of controversy. He notes, as Scanlon also does in his commentary, that although Parfit aims to show that the most philosophically profound aspects of Kant's moral theory can be defended as a pillar of the Triple Theory, Parfit's Kantian Contractualism is deeply un-Kantian in taking rationality to depend mostly on independent facts about normative reasons, rather than vice versa. It is crucial to Parfit's Convergence Argument that these reasons prominently include reasons of impersonal good "in the impartial-reason-involving sense" (1:41). A state of affairs is good in this sense if everyone has a normative reason to prefer that it obtain. Obviously, this way of viewing the relationship between practical rationality and normative reasons will seem uncongenial to many Kantians and to reorient their favored approach toward consequentialism from the outset.

Volume 1 is divided into three "parts," which Parfit respectively titles "Reasons," "Principles," and "Theories." However, these do not really reflect the structure of the volume's project, which has three main elements: (i) conceptual preliminaries and substantive issues about normative reasons in general, (ii) arguments that the Triple Theory's constituent principles best express the most profound insights of Kantianism, Contractualism, and Consequentialism, respectively, and (iii) the Convergence Argument. Part 1 does focus on (i). But element (ii) extends throughout part 2 ("Principles") and much of part 3 ("Theories"), with Kant's theory receiving by far the most discussion: all of part 2 and the first three chapters of part 3 (1:177–342). Only one chapter each are devoted to Contractualism and Consequentialism (1:343–403). Much of the Convergence Argument is implicit in (ii), so much, indeed, that its formal statement only requires fifteen pages (1:404–19).

Part 1 ("Reasons") is mostly concerned with the concept and normative theories of normative reasons, mostly reasons to desire and act, but also epistemic reasons for belief, and, in principle, reasons for any attitude for which there can be reasons. Parfit's main goal is to argue that normative reasons for attitudes, like desire and belief, are invariably *object* *given* rather than *subject or state given.* By this, he means that, for example, all reasons for desire consist in (object-given) facts about possible objects of desire rather than any (state-given) fact about the attitude of desire itself, for instance, that having the desire would have good consequences or help to constitute an intrinsically good state of affairs. Parfit then argues on this basis against *Subjectivist Theories* and for *Objectivist Theories* of normative reasons for desire and action. Parfit also calls objective theories *value-based theories* since they hold that all normative desiderative and practical reasons are (object-given) good- or value-making facts.

Subjectivist theories hold that normative reasons are given by, or depend upon, some fact about the agent's desires, for example, that the act would fulfill a desire the agent has or would have after "procedurally rational" "ideal deliberation" (1:64). Parfit defines Objectivism about Reasons in such a way that it simply follows from the thesis that all reasons for acting are object given rather than state given (1:45). But this is at best misleading since it is possible that all normative reasons are object given, but also that the existence of object-given reasons itself depends upon certain state-given facts, for example, that awareness of the object-given facts would under conditions of ideal deliberation motivate the agent to have desires the act would fulfill (see, for example, Darwall 1983, 78–82; Schroeder 2010, 23–40). I shall argue that this is the most plausible version of Subjectivism (or "existence internalism" about reasons) and that it can escape some (though not all) of the arguments that Parfit lodges against it.

The last two chapters of "Reasons" concern morality, and each does important spadework for the Triple Theory and for the Convergence Argument that crowns volume 1. Chapter 6 ("Morality") begins with Sidgwick's "dualism of practical reason," which Parfit reformulates as a conflict between *Rational Egoism* and *Rational Impartialism* ("We always have most reason to do whatever would be impartially best") (1:130).[8] A crucial linchpin in the Convergence Argument is what Parfit calls a "wide value-based objective view" of practical reason, which strikes a compromise between Rational Egoism (perhaps better, "Rational Partialism") and Rational Impartialism, holding that when reasons of partial and impartial good conflict, "we often have sufficient reasons to act in either of these ways" (1:137). In other words, in many such cases neither partial

[8] The difference is that Sidgwick's impartial principle is a principle of rational *benevolence* according to which the agent has reason to pursue "the good of any other individual as much as his own" (Sidgwick 1967, 382).

nor impartial reasons are decisive; we may act as either recommend and not be doing something we have most reason not to do.

Chapter 7 ("Moral Concepts") concerns the distinctive character of the deontic "group of concepts": moral right and wrong, obligation, requirement, permission, and so forth, that feature in the Triple Theory (1:165). This chapter bears special scrutiny since it reveals grounds for conceptual constraints that push toward a Rawlsian publicity condition and, therefore, away from an act consequentialist theory of right. Indeed, Parfit here observes that act consequentialism, at least in an "impartial reasons" version that seems similar to the view he defended in *RP*, "may be better regarded, not as a moral view" at all, but "as being, like Rational Egoism, an external rival to morality" (1:168).

Part 2 marks the beginning of *OWM*'s engagement with Kant. Part 2's main focus is Kant's "best-loved principle," the Formula of Humanity (FH), in Parfit's formulation: "We must treat all rational beings, or persons, never merely as means, but always as ends" (1:177). Parfit here discusses the ideas of consent, the dignity of persons, and Kantian issues about treating persons as ends and means, mostly with an eye to arguing that these supply no defensible principle that is independent of the Kantian Contractualist Formula. Chapter 10 ("Respect and Value") makes some points that will have special importance to my claims below, namely, that there are forms of *value*, specifically those instantiated by persons and, indeed, by morality itself, that are not, or at least not just, kinds of *goodness.* Unlike the good, which is analytically related to normative reasons for a *desire* that something exist or obtain, the dignity of persons is connected conceptually to reasons for the very different attitude of *respect.* Value of this distinctive sort is connected to deontic concepts analytically.

Parfit then turns in the first half of part 3 (chapters 12–14) to Kant's Formula of Universal Law and to related notions of impartiality, reversibility, and the Golden Rule. Here he argues that a defensible principle of right can be grounded in these ideas, namely, the Kantian Contractualist Formula: "Everyone ought to follow the principles whose universal acceptance everyone could rationally will" (1:342).

Chapter 15 focuses on Contractualism and chapter 16 on Consequentialism. The former includes a relatively short critical discussion of Rawls's version and elaboration of Scanlon's, ending in the formulation that will be adapted for the Triple Theory: "Scanlon's Formula: An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject" (1:369).

Chapter 16 defines Consequentialism sufficiently broadly to include both Act and Rule forms.

> *Consequentialism*: Whether our acts are right or wrong depends only on facts about how it would be best for things to go (1:373).

'Best' means "in the *impartial-reason-implying sense*" (1:371). Something is good in this sense if there is, from an impartial point of view, reason for everyone to want that thing. So an outcome of a possible act is best in this sense if there is *more* reason, or at least, not less, for everyone impartially to want that outcome rather than that of any other available act.[9]

Much of chapter 16 is taken up with a "Kantian Argument" for Rule Consequentialism. Parfit begins with the Kantian Contractualist Formula and argues, roughly, that since what is impartially best is what everyone has impartial reasons to desire, it follows that everyone has impartial reasons to will that everyone accept principles, the universal acceptance of which would make things go impartially best. So far, this is pretty much a conceptual truth. A wide value-based objective view of practical reasons then ensures that these pro tanto reasons will be sufficient reasons. Furthermore, Parfit argues, there are no other principles that *everyone* has sufficient reason to will that everyone accept.

Chapter 17 completes the Convergence Argument. Parfit argues first that any principle that satisfies the Kantian Contractualist Formula satisfies Scanlonian Contractualism ("Everyone ought to follow the principles that no one could reasonably reject"). The leading idea here is that if a principle is such that everyone has sufficient reason to will everyone's acceptance of that principle in preference to alternatives, then "no one's objection to this principle could be as strong as the strongest objections to every alternative" and therefore no one could reasonably reject it (1:411–12). Since we already have that any principle that satisfies the Kantian Contractualist Formula also satisfies Rule Consequentialism, we can now conclude that any principle satisfying either of these will also satisfy Scanlonian Contractualism. Volume 1 ends with the summit of the Triple Theory having been achieved.

[9] This value is *impersonal*; impartial (or impersonal) goodness is not, like that of well-being or benefit, goodness *for* anyone in particular, even when the impartially good-making feature of the outcome consists in the realization of something good for, or that benefits, someone. What matters is the further fact that the existence of this benefit is a good thing in itself (impartially or impersonally considered).

Like volume 1, volume 2 falls into three "parts." Part 4 consists of commentaries on volume 1 by Susan Wolf, Allen Wood, Barbara Herman, and Scanlon, which Parfit responds to in part 5. Parts 4 and 5 are filled with interesting points, but the most substantial is Parfit's extended exchange with Scanlon, which has special relevance for Parfit's claims about the Triple Theory. Scanlon and Parfit both make telling points to which we shall return below: Scanlon about how Contractualism might resist any implication of Rule Consequentialism, and Parfit about how Contractualism might more plausibly deal with issues concerning aggregation and "how the numbers count" (the topic of chapter 21).

By far the most significant part of volume 2 is part 6 ("Normativity," 2:263–620), which focuses on metaethics. For Parfit, the stakes here are especially high since unlike many philosophers, Parfit holds that the results of normative theory depend on metaethics. The Triple Theory purports to tell us what matters morally, at least, so far as our obligations are concerned. But Parfit believes that whether anything matters at all depends on the truth of a metaethical theory according to which normative claims, facts, and properties are irreducibly normative and nonnatural, "Meta-ethical Cognitivism," as he calls it.

Part 6 begins with extended critiques of alternatives to metaethical cognitivism: naturalism in all its forms (2:263–377), expressivism, and other forms of antirealism (2:378–463). Putatively normative claims can be robustly true according to naturalism, but they end up not being genuinely normative; normativity and mattering themselves never come sufficiently into view. Expressivists succeed better in focusing on the normative, but they do not hold that normative claims can be true "in the strongest sense." On neither view, Parfit concludes, can genuinely normative claims be true in this sense.

A notable absence from Parfit's canvassing of alternatives to Non-naturalist Cognitivism is any form of constructivism that is not committed to metaethical naturalism. Scanlon points out that "Kantian constructivism" might seek to ground facts about normative practical reasons in claims about what a rational agent is committed to in practical reasoning and in "seeing herself as a rational agent" (2:121). He notes that Parfit's version of Kantian Contractualism appeals to "true substantive claims about reasons" that hold independently of anything presupposed in practical reasoning and, we might add, of the deliverances of "procedurally rational" deliberation (2:123). Parfit takes note of these points with breathtaking brevity. To the question: Why ought we to reject Kantian constructivism "and appeal instead to . . . 'true substantive claims about reasons'?" Parfit answers, "We ought to appeal to such claims, I believe, because they are true" (2:191). And he adds that for Kantian moral theories to generate plausible results, they must go beyond claims about we can rationally will consistently "with regarding ourselves as rational agents." Such claims are, he says, "too restricted, and too weak" (2:191).[10]

Parfit terms his own metanormative position "Non-Metaphysical Cognitivism" since although it holds that normative concepts, propositions, truths, and properties are all irreducible and nonnatural, it also maintains that normative truths have "no positive ontological implications" (2:479). This might sound like a "quasi-realist" expressivism of the sort associated with Allan Gibbard and Simon Blackburn, which Parfit argues are no less Noncognitivist than the expressivisms of A. J. Ayer or R. M. Hare. But Parfit insists it is not. Unlike quasi-realist expressivists who embrace a minimalist account of normative truth, Parfit claims that normative claims can be true "in the strongest sense" (2:479). His model is mathematics, of which he takes a similarly nonmetaphysical cognitivist view. Whether numbers exist in some ontological sense is, Parfit claims, not relevant to whether, say, there are one thousand consecutive zeros in the decimal expansion of pi (2:481). And the same is true, Parfit thinks, when it comes to normative truths.[11]

Parfit takes on epistemological challenges to Nonnaturalist Cognitivism more directly. Chapter 32 (2:488–510) is concerned with the "Causal Objection: Since non-natural normative properties or truths could not have any effects, we could not have any way of knowing them" (2:488). Chapter 33 (2:511–42) concerns the related "Darwinian Dilemma" that Sharon Street (2006) has argued is especially pressing for nonnaturalist "realist" views like Parfit's.[12] On the assumption that natu-

[10] In Darwall 1983, 1990, and 1992, I defended what I thought of as a kind of Kantian internalist or constructivist position. Parfit discusses some of my views in his critique of internalist and naturalist approaches, which I will comment on below.

[11] Because he denies that normative truths have ontological or metaphysical implications, Parfit prefers not to refer to his view as a form of "realism." The entry for "normative realism" in the index reads as follows: "belief in normative truths that are not response-dependent, mind-dependent, or constructivist; often assumed to make positive ontological claims; the word 'realism' not used by me for this reason" (2:823). Parfit calls his view Nonnaturalist Cognitivism. Beyond, however, insisting that normative facts and properties are irreducibly nonnatural and that normative claims can be true "in the strongest sense," there seems not much more that Parfit can say about the metaphysics (or nonmetaphysics) of the normative.

[12] Street also raises the worry for some versions of naturalism. "Realism" is her term. Parfit again prefers "cognitivism" for his "nonmetaphysical" metaethical view.

ral selection has played a significant role in shaping human attitudes and normative beliefs, a nonnaturalist metaethics must either (implausibly) hold that this process has somehow involved the nonnatural normative facts or explain how our normative beliefs could even be in the neighborhood of these facts despite this form of natural influence.

Against the Causal Objection, Parfit argues that normative knowledge, like mathematical knowledge and unlike ordinary empirical knowledge, does not require any causal responsiveness to the objects of knowledge. The epistemology of Nonmetaphysical Cognitivism is "*Intuitionism*," the view that "we have *intuitive* abilities to respond to reasons and to recognize some normative truths" (2:544), combined with the method of reflective equilibrium. Criticism of intuitionism generally stems, Parfit thinks, from the "view that intuition is a special quasi-perceptual faculty," but Parfit rejects this picture (2:544). "When I use the word 'intuitive,'" he says, he means what even critics do when they say that a claim is "intuitively plausible" or "intuitively clear" (2:545).

But Nonmetaphysical Cognitivism abjures any *explanation* of how our intuitions manage to cotton on to normative truth or how we manage to detect and respond to normative reasons. It can provide no metaphysical account of what normative properties and facts are, other than that they are irreducibly normative and nonnatural, to help us understand how we might come to know them. If calling something "intuitive" just means that it is "intuitively plausible," all we have is that it can strongly *seem* to us as if some normative claim we are considering is true.

From this perspective, the worry is that we lack any account of what normative properties and facts are that might help explain how our apparent responses to normative reasons or normative intuitions can amount to knowledge. But there are also worries from, as it were, the opposite perspective. As Street's challenge makes clear, some, perhaps many, of our intuitions can be explained in some other way, for example, as the result of evolutionary influences. Parfit takes up this challenge, arguing that our normative intuitions and beliefs are not plausibly regarded as resulting from evolutionary pressures (2:511–42). He grants that there may be good evolutionary explanations of our responsiveness to epistemic reasons (for example, forming beliefs on the basis of inductive evidence), but this does not mean that epistemic beliefs about normative reasons can be so explained (2:515). And even if evolution can explain various "strong motivations" human beings have, these do not explain any human responsiveness to normative practical reasons, much less any normative practical beliefs (2:528).

Granting all of this, however, there still seems to be a way of raising Street's worry, namely, that what Parfit calls "motivations" may be difficult if not impossible to distinguish phenomenologically from apparent responsiveness to reasons. "Responding to reasons," as Parfit uses it, is a success notion, but we can presumably speak of its subjective aspect: seeing something as a reason or its seeming to one to be a reason. And how does this differ from motivations, viewed subjectively? If desires involve, as Scanlon supposes, being disposed to see features of its object as normative practical reasons, what can provide a bright line between motivations and practical normative intuitions (Scanlon 1998, 39)?

Parfit recognizes that the possibility of disagreement can challenge Nonmetaphysical Cognitivist Intuitionism even beyond the general Sidgwickian epistemological worry mentioned earlier. In the latter case, the issue is not so much whether truths exist in some relevant domain, but who has access to it. But if there is nothing to explain how or why normative intuitions correctly represent the normative facts when they do, and if intuitions do not converge in reflective equilibrium, then there seems reason to doubt that there exist normative facts and properties for intuitions to respond to. So Parfit argues at length (2:543–606) for the "Convergence Claim," roughly, that procedurally ideal reflection on all the nonnormative facts leads nearly everyone to sufficiently similar normative beliefs.

Both volumes 1 and 2 are thus completed by distinct convergence claims. Volume 1's Convergence Argument argues that three central moral philosophical traditions converge on the Triple Theory, telling us what matters morally in the deontic sense. Volume 2's Convergence Claim, by contrast, defends against the skeptical worry that there might be no normative facts and hence, that nothing matters, whether morally or otherwise.

In the space remaining, I shall concentrate on four aspects of Parfit's overall argument: (i) his claims about normative reasons, (ii) his discussion of the deontic concepts of moral right and wrong, (iii) his arguments for the Triple Theory, and (iv) his arguments against all forms of ethical naturalism.

Normative Reasons. Human beings are, Parfit says, "the animals who can both understand and respond to reasons" (1:31). But what is the relation between the capacity to respond to reasons, on the one hand, and the existence of normative reasons, on the other? Existence internalists or "Subjectivists," as Parfit calls them, suppose that there is a very intimate relation.[13] They believe that in order for a normative reason to exist for agents (construed broadly, to include epistemic agents, desiderative agents, and so forth), it must in principle be possible for them to act or form the relevant attitude *for that reason*—that is, for it to be *their reason* for doing so—and that they would do so were they to exercise ideally the powers of deliberative reasoning that are essential to them being agents of the relevant kind. Objectivists hold, by contrast, that normative reasons do not depend on any such "subjective" counterfactual. Where subjectivists (existence internalists) see normative reasons as essentially things that agents can take account of in deliberation, from the agent's point of view, objectivists see their existence as being independent of any such deliberative perspective and reasoning. Objectivists of course agree that normative reasons must hook up with deliberative reasoning in the sense that were agents *substantively rational*, they would take up normative reasons in their deliberative reasoning, be moved by them, and so make them *their* reasons (1:73). But that just means that they would do so if they were to respond to the normative reasons that exist for them independently.

Parfit focuses on normative reasons because he takes all interesting normative questions and concepts to be "reason implying." Roughly speaking, concepts are normative in a "reason-implying sense" when their instantiation analytically entails the existence of normative reasons for some attitude or other (understood broadly to include intention and action). Normative epistemic concepts concern reasons for belief, normative practical concepts concern reasons for action, and so on. Some of the latter concern goodness or desirability, that is, what there is normative reason *to desire* (whether impartial goodness in the "impartial reason-implying sense," or forms of partial or personal desirability, where the reasons exist from more partial perspectives). Parfit does not believe these are the only normative concepts bearing on action. As we noted above, there are also deontic concepts (though, as we will see, their relation to deontic practical reasons is a complicated matter). And there are forms of value, like the dignity of persons, that entail normative practical reasons, however, reasons to respect rather than to desire and promote.

In any case, normative reasons are where the action is, both in normative theory and in the metaethics of normativity. And much of

[13] For the distinction between existence internalism and judgment internalism about reasons, see Darwall 1983, 54–55, 81; and Darwall 1997.

Parfit's discussion focuses on combating Subjectivism in various forms, whether in substantive normative or in metaethical versions (1:43–110; 2:263–94). We can usefully distinguish three main versions (focusing on their most plausible formulations and on normative *practical* reasons).

Analytical Subjectivism: Sentences like "There is normative reason for S to do A" mean the same as "A would best fulfill the desires S would have were he or she fully informed and to deliberate in an ideally (procedurally) rational way" (1:72).

Substantive (Normative) Subjectivism: There is normative reason for someone to do something only if that action would satisfy a desire he or she actually has or would have were he or she fully informed and to deliberate in an ideally (procedurally) rational way.

Metaethical Subjectivism (Weak Form): There can be normative reason for someone to do something only if it would satisfy the desires he or she would have were he or she fully informed and to deliberate in an ideally (procedurally) rational way.

Metaethical Subjectivism (Strong Form): There being normative reason for someone to do something consists in its being the case that the action would satisfy the desires he or she would have were he or she fully informed and to deliberate in an ideally (procedurally) rational way.[14]

Note that these formulations are neutral between naturalist (broadly Humean) versions, like Williams's, and Kantian versions, like Korsgaard's, which hold that rational agents are committed to substantive moral ends by presuppositions of the deliberative standpoint. The central point, as Parfit emphasizes, is that Subjectivists (existence internalists) hold that normative reasons for action are constrained by what the agent could or would *take* as reasons when *procedurally* rational. Substantive rationality neither constrains nor provides any guide to normative reasons' metaphysical character.

Parfit convincingly argues against Analytical Subjectivism that it makes putatively normative claims into "concealed tautologies." When we say that someone ought or has reason to do what will satisfy his or her informed procedurally rational desires, we end up just saying that this will satisfy those desires rather than anything with normative force. Parfit takes Bernard Williams to be an Analytical Subjectivist because Williams

[14] I am deliberately ignoring "conditional fallacy" issues of a sort that have led to more sophisticated formulations, for example, in Railton 1986 and Smith 1994, in terms of what an ideally deliberating version of the agent would want for himself or herself in his or her actual nonideal circumstances (see also Johnson 1999).

sometimes speaks of reasons in an "internal sense." So Parfit argues that he and Williams have no real disagreements about normative reasons because Williams isn't really talking about them at all. (Parfit takes the concept of a normative reason to be basic and irreducible.) While this is not an entirely uncharitable interpretation of what Williams says, we get a more plausible view when we read Williams and other existence internalists as holding either weak or strong forms of Metaphysical Subjectivism.

Parfit's discussion of Substantive Subjectivism focuses on his arguments that reasons must be object given rather than state given together with various intuitive arguments (for example, that we have reason to avoid future agony independently of any subjectivist fact) (1:73–82) and arguments that more sophisticated "procedurally rationalist" forms of Subjectivism, such as those defined above, are actually incoherent (1:83–101). However, it can be argued that Metaethical Subjectivism, whether weak or strong, is actually especially well suited to explain why reasons must be object given rather than state given.

If reasons are, in their nature, considerations that agents can *see* as reasons from the perspective they take up in deliberating about what to desire, believe, and do, then we need to ask what things look like *from that perspective.* Deliberating agents' attention is not on their own subjective states, but on these states' objects. Agents' desires and beliefs are "backgrounded" (Pettit and Smith 1990; Darwall 1983, 37–42). What they take as reasons and are motivated by are not that they have the desires and beliefs they do (even as the result of ideal deliberation), but object-given facts they see from the perspective their desires and beliefs give them. To be motivated by a desire not to die, for example, is to be motivated by facts about dying, not by facts about a desire not to die.

Of course, some state-given facts might also be object given. An agent might think about the fact that he or she wants something and be motivated by this fact even beyond any (other) object-given facts. ("You know what? I just want to watch that trashy sitcom, and damn it, I'm going to do it.") But here the consequent second-order desire, to do what he or she wants to do, is not his or her reason. From the perspective of the second-order desire, the first-order desire seems an object-given reason, and the state-given fact of the second-order desire seems no reason at all (Darwall 2001).

Thus, far from denying that reasons are object given rather than state given, the Nonanalytical Metaethical Subjectivisms defined above seem committed to the *opposite* position. What's more, they can explain why normative reasons are object and not state given in a satisfying way.

Parfit says that "if we believe that all practical reasons are [object given], we are Objectivists about Reasons" (1:45). But that is not so if Objectivism is supposed to contrast with the forms of Subjectivism mentioned above, as Parfit obviously intends. These views distinguish between the object-given facts that are normative reasons, and a subjective counterfactual condition that is necessary for these object-given facts actually to be normative reasons for an agent (see, for example, Darwall 1983, 78–82; Schroeder 2010, 23–40). Any Subjectivism worth bothering with is best understood as a metaethical view (Metaethical Subjectivism), albeit one with normative consequences (Substantive [Normative] Subjectivism).

Parfit argues that Substantive Subjectivism is committed to unintuitive normative consequences, for example, that an agent might have no reason to avoid future agony. Whether that is so depends on the details of ideal procedural rationality. It seems likely that some versions, perhaps Williams's, will face this objection, but it is more difficult to assess the degree to which any version must be. But it is certainly right that no version of Subjectivism can say that an agent has reason to avoid future agony *simply* because of what agony is. And stipulating that deliberative rationality itself includes equal concern for oneself at all times is gerrymandering.

More potentially damaging is Parfit's claim that procedural rationalist versions of Subjectivism are actually less coherent than less sophisticated forms, like Williams's "sub-Humean model," which focus on agents' actual desires. Why think that agents should seek further information unless this concerns reasons for acting that hold independently? However, more sophisticated forms of Metaethical Subjectivism can answer this question. When we get more nonnormative information, this can lead to rational revision in our desires and consequent reason takings. Believing that it would give a child pleasure, I want to give her a toy and see the object-given fact of her pleasure as a reason to do so. Learning that her pleasure will consist in tormenting her brother with the toy, however, I want, for this reason, not to give it to her.[15] The latter is a more inclusive

[15] I don't have to learn a further normative fact or make a further normative judgment that the fact that she will get pleasure is not a reason, or less of a reason, since this pleasure will be at her brother's expense. I assume that there is a form of taking or seeing something as a reason consisting in being motivated by it that is more primitive than, and does not depend on, a normative *judgment.* Think, for example, of the way emotions like the fear of flying involve reason takings in the former sense even in the face of opposing normative judgments.

view; it takes account of the fact of the one child's pleasure and the fact that this pleasure will itself have as its object her brother's pain. The Metaethical Subjectivist can hold that deliberation based on more inclusive knowledge is more rational or deliberatively ideal even when this knowledge is not itself explicitly normative (that is, concerning normative reasons as such).

Wrongness. Volume 1's crowning Convergence Argument concerns moral right and wrong. I noted earlier that all three of the Triple Theory's component principles respect a Rawlsian "publicity condition." They judge the wrongness of acts by assessing what things would be like were everyone to accept in common some principle (P) that would disallow the act. Call this the "P world." Kantian Contractualists look to whether everyone can rationally will the P world. Rule Consequentialists look to whether impartial goodness of the P world is greater than that of any alternative "principle world." Scanlonian Contractualists look to whether no one could reasonably reject the P world.

I suggested above that what pulls Parfit's thought in the direction of these theories and away from the sort of consequentialism he defended in *RP* is that he now believes that moral right and wrong are distinctive deontic concepts that differ from that of the good in important respects. Chapter 7 has a very careful discussion of different senses of moral 'wrong'. Prescinding from differences between "fact-relative," "belief-relative," and "evidence-relative" senses, we can focus on those Parfit calls, respectively:

> *Blameworthiness sense*: 'Wrong' means 'blameworthy' (1:165).
> *Reactive attitude sense*: 'Wrong' means 'an act of a kind that gives its agent reasons to feel remorse or guilt, and gives others reasons for indignation and resentment' (1:165).
> *Decisive-moral-reason sense*: 'What we ought to do' means 'what we have decisive moral reasons to do' (1:166).
> *Morally-decisive-reason sense*: 'Wrong' means 'what we have morally decisive reasons not to do' (1:167).

In addition, there is what Parfit calls the "ordinary" sense of 'morally wrong', which he says is indefinable but whose sense can also be expressed by the phrase 'mustn't be done' (1:165).

In Darwall 2006, I argue that deontic moral concepts involve the reactive attitude sense. Since reactive attitudes, within which I include moral blame, have an ineliminably second-personal character (they are implicitly addressed to their objects), I conclude that these concepts are second personal in this sense (see also Strawson 1968, Watson 1987, and Wallace 1994). The idea is not that 'wrong' means the same as 'blameworthy' since someone may do something wrong without being to blame if he or she has an excuse. I claim, rather, that it is a conceptual truth that an action is wrong if and only if it is an action of a kind such that the agent would be blameworthy (or justifiably the object of guilt, resentment, or indignation) were he or she to perform it without excuse.

Parfit distinguishes between the "ordinary" and reactive attitude senses, but he sometimes relies on the latter, for example, when he argues that Sidgwick's act consequentialist principle of universal benevolence may be better regarded as an "external rival to morality" (1:168). "When Sidgwick claims that he *ought* not to prefer his own lesser good [to greater impartial good], he does not seem to mean that such a preference would be blameworthy . . . or that such an act would give him reasons for remorse and give others reasons for indignation" (1:168). And Parfit argues against the decisive-moral-reasons sense that the claim that we might sometimes have sufficient reason to do what is morally wrong is not a concealed contradiction (1:167).

There is another reason for rejecting this sense, namely, that the fact that an action is wrong can itself provide a reason not to perform it (a deontic reason as Parfit calls it) (1:172–74; see also Darwall 2010). That would not be so if being wrong meant just that there exist (other) morally decisive reasons not to perform the action.

But what about the morally-decisive-reason sense? Is it a (perhaps concealed) contradiction to claim that it might not be morally wrong to do what there is no morally sufficient reason to do? That depends on what it is for a reason to be morally sufficient and not outweighed morally. Consider a case where a great good can be accomplished only at great cost to the agent—say rescuing someone from a burning building. It seems that we can coherently think that the costs to the agent might make not attempting the rescue not morally wrong, although attempting it might remain morally better (if the risks are not so great as to make the rescue foolhardy). In other words, it seems perfectly coherent to think that such a rescue would be supererogatory. But if we think supererogation is a coherent possibility, we cannot be using 'wrong' in the morally-decisive-reasons sense.

This is perfectly understandable if we use 'wrong' in the reactive attitude sense. If someone thinks we cannot reasonably demand that agents assume such a cost, then he or she will think that it is not morally wrong for them not to do so. As Strawson and his followers have pointed out, to hold someone responsible with reactive attitudes like moral blame *is* implicitly to make a demand of them (Strawson 1968, Watson 1987, Wallace 1994, Darwall 2006). So our thought is completely intelligible if by 'wrong' we mean 'action of a kind that would be blameworthy and justifiably the object of reactive attitudes like guilt, were it done without excuse'.

There is also a significant problem with thinking that the "ordinary" sense of 'wrong' expresses a concept that is irreducible in the same way that the concept of a normative reason is, namely, that the concept of moral wrong will then cease to be a normative concept. If what makes a concept normative is that it is normative reason implying, then it would seem that normative reasons must somehow enter into its analysis.[16]

I suggest that the ordinary sense of 'wrong' *is* the reactive attitude sense. And this explains why and how 'wrong' expresses a normative concept while, unlike the decisive-moral-reasons sense, leaving it open that there might sometimes be sufficient reason to do moral wrong and, unlike the morally-decisive-reasons sense, leaving open the possibility of supererogation. On the reactive attitude sense, an action's being wrong does analytically entail that if the action is done without excuse, then there are normative reasons for blame (and other reactive attitudes) if the action is done without excuse. Taking the ordinary sense to be the reactive attitude sense thus gives us an explanation of why 'wrong' expresses a normative concept. But if we take the concept to be irreducible in the same way that the concept of normative reason is, we can no longer take the concept to be normative in a normative-reason-implying sense.

It is clear in any case that Parfit takes the ordinary sense of 'wrong' to be close to the reactive attitude sense. And once we see this, we can see why he is led in the direction of a Rawlsian publicity condition as a "constraint of the concept of right." So much follows from a Strawsonian view of what it is to hold someone accountable (someone else *or* oneself) with a reactive attitude like moral blame. We have such attitudes from a distinctively interpersonal (or second-personal) perspective in which we presuppose that the standards to which we hold one another are available to everyone in common (Strawson 1968, Darwall 2006). We cannot intelligibly hold someone accountable for complying with an inaccessible esoteric principle. If moral right and wrong are tied to accountability

[16] Nor does it to help to fix the concept of moral wrong to say that it can also be expressed by the phrase "mustn't be done." A doctor can convincingly say that a patient must not smoke without implying that it would be morally wrong for him to do so.

conceptually, any theory whose (putative) justifiability depends on remaining esoteric will seem "an external rival to morality" (1:168).

The Convergence Argument. Moreover, once one accepts that deontic concepts like moral wrong are tied to accountability conceptually, grounds emerge to resist Parfit's Convergence Argument, both (i) from Scanlonian Contractualism to Rule Consequentialism, and (ii) from Kantian Contractualism to Rule Consequentialism. And these reasons also support (iii) preferring Kantian and Scanlonian Contractualism over Consequentialism, even in its rule form.

According to Scanlon's "contractualist ideal," we realize "mutual recognition" or reciprocal respect when we live in ways we can justify *to* one another and hold ourselves mutually accountable for complying with principles that no one can reasonably reject (Scanlon 1998, 162). If the form, as it were, of contractualist principles is to mediate mutual accountability, then to consider whether a principle can reasonably be rejected, we must ask whether anyone can reasonably reject being held accountable for complying with it, that is, their being subject to a demand to do so.

Parfit's argument from Scanlonian Contractualism to Rule Consequentialism depends on his prior argument that any principle that satisfies Kantian Contractualism will also satisfy Rule Consequentialism and vice versa (1:411–12, 377–400). The basic idea is that the notion of good that enters into any form of Consequentialism is that of impartial good in the reason-implying sense.[17] Whether an outcome is good in this sense is just the question of whether there is normative reason for anyone to want it from an impartial point of view. Kantian Contractualism holds, roughly, that it is wrong not to comply with principles that everyone has sufficient reason to will to be universally accepted (as principles of moral right). Call, again, the world in which some principle P is thus universally accepted, the P world. According to Kantian Contractualism, then, it is morally obligatory to comply with P, and wrong to violate P, just in case everyone has sufficient reason to will the P world in preference to any Q world, where Q is any possible alternative principle for a situation of the relevant kind.

[17] Of course, the most familiar forms of Consequentialism, Utilitarianism, and other forms of Welfarism also involve the idea of someone's good, benefit, or well-being, but even here, they hold that the impartial goodness of people benefiting figures essentially into what makes actions morally right or wrong.

Let us now say that principle P and the P world are "optimific" if there is more impartial good in the P world than in any Q world. It simply follows from the definition of impartial good that all have pro tanto impartial reason to want, hence will, optimific principles and worlds. And it follows further from Parfit's "wide value-based objective" view of reasons that these pro tanto impartial reasons can be *sufficient* reasons, that is, not decisively overridden. Of course, for every individual, there will likely be *some* principle and world that he or she has reason to will other than, or perhaps in preference to, optimific ones. But there are arguably no other principles that *everyone* has reason to want or will in preference to optimific ones. It apparently follows that according to Kantian Contractualism, it is wrong to violate optimific principles. Parfit concludes that Kantian Contractualism entails Rule Consequentialism.

Presently, I shall suggest a problem with this argument. First, however, consider how Parfit extends it to Scanlonian Contractualism. Seeing how a Scanlonian can resist the argument may help us see how a Kantian Contractualist can also. Suppose, however, that the argument from Kantian Contractualism to Rule Consequentialism goes through. Assume, that is, that the only principles that everyone can will everyone to accept are optimific principles. When that is so, Parfit says, "there must be facts which give us a strong objection to" any alternative principle, and so no one's objection to optimific principles can be as strong (1:411). So when this is so, he argues, no one can reasonably reject the optimific principles. It follows that principles that satisfy Rule Consequentialism will also satisfy Scanlonian Contractualism.

The problem with this argument is that it treats whether someone has a *reasonable objection* to a principle P to turn on whether the reason he or she might give outweighs other reasons *as reasons to desire (and therefore will) the P world.* Since, by hypothesis, optimific principles are those that everyone has reason to want, therefore will, that everyone accept, any objection that someone might make to some principle P, *insofar as it concerns a reason not to want the P world,* has already been weighed in, and, by hypothesis, been outweighed. So if the optimific principles are the only ones everyone has reason to will that everyone accept, then it could not be the case that anyone has a reasonable objection to optimific principles. The possibility will have been ruled out.

But this is not, I think, the best way to understand the idea of reasonable rejection or objection within Scanlonian Contractualism. Grounds for reasonable objection and rejection derive not fundamentally from reasons for wanting certain outcomes, but from considerations that can ground *claims* and demands that can reasonably be made on people. It is entirely compatible with someone having sufficient reason to want a "principle world" because of the impartial good that would result, that, at the same time, he or she could reasonably object to and reject the imposition of the demand that such a principle would represent. Whether the imposition of a demand on someone is justified (or whether that person could reasonably object to it) depends on reasons or grounds that can support or defeat *claims* or *demands.* This is evidently an issue of a different kind than whether there might be good reasons, of whatever kind, for wanting certain outcomes. So even if the fact that the P world is optimific gives everyone a reason to want or will that everyone accept P, that does not show that there might not be reasonable objections to the demand that P would represent since the reasons for willing the P world might not be reasons of the right kind to support the imposition of a demand. There might be reasonable objections to such a demand even if everyone could rationally desire and will the P world.

As I see it, the fundamental idea underlying contractualisms of the sort advanced by Scanlon and Rawls is that, as Rawls put it, persons are "self-originating source[s] of valid claims" (Rawls 1980, 546). Everyone has the same basic authority to make claims and demands of one another as everyone else, and principles of right are those no one could reasonably reject (from a position of having this equal basic authority) as standards to which all are held mutually accountable.[18] Parfit's argument simply assumes that any impartial reason for wanting outcomes that would result from people being subject to a demand will automatically support the imposition of the demand without any showing of how it relates to anything anyone can validly claim or that can support valid claims. It is consistent with a reason's supporting someone's freely accepting a burden necessary to contribute to impartial good or prevent impartial bad that that reason might nonetheless not support any claim or demand that he or she accept the burden. The burden might, consistently with the impartial goods at stake, be a burden he or she could reasonably reject since it was not sufficiently grounded in anything that could support a legitimate claim or demand.

There is thus a conceptual gap between reasons for desiring certain outcomes, even impartially, and reasons—second-personal reasons, as I call them—that concern valid claims and demands. Parfit himself

[18] In my work, this is what I call equal basic second-personal authority (Darwall 2006; see esp. 300–320).

notes this difference in passing when he criticizes what he calls Scanlon's "individualist restriction," according to which "numbers do not count" when many stand to benefit or be harmed. In considering whether someone can reasonably reject a principle, according to Scanlon, we must consider what is at stake for that person and other *single* people (2:193). Parfit plausibly notes that Scanlon could avoid consequentialist conclusions while allowing numbers to count if he were to hold what Parfit calls the "Contractualist Priority View: People have stronger moral claims, and stronger grounds to reject some moral principle, the worse off these people are" (2:201). Parfit himself observes that this is *not* a view "about the goodness of outcomes" (what he calls the "Telic Priority View"). But if that is right, facts about impartial good are distinct from those concerning moral *claims* and cannot, without some further account of how they can ground such claims, show that rejecting a principle would be *unreasonable.* It follows that the Convergence Argument from Scanlonian Contractualism to Rule Consequentialism does not go through.[19]

Moreover, this points to a way of resisting Parfit's argument from Kantian Contractualism to Rule Consequentialism, namely, by interpreting Kant's idea of the dignity every person has as an end in Rawlsian terms as a "self-originating source of valid claims." As Parfit himself recognizes, the distinctive value that Kantians believe persons have differs from any kind of good, including any impartial good that might be desired or promoted; it is rather to be respected (1:233–50).[20] It will follow from a contractualist interpretation along Rawls/Scanlon lines that to decide which principles we can rationally will that everyone accept, we have to ask what principles could be justified from the perspective of an equal basic authority to make claims and demands of one another. And again, if that is right, no simple argument from the definition of impartial good and a

[19] In response to Scanlon's raising a version of the problem I discuss here for the Convergence Argument (2:135–38), Parfit notes that *if* someone can reasonably reject a particularly burdensome optimific principle that he or she also has sufficient reason that everyone accept, then "Scanlonian Contractualism would here conflict with Kantian Rule Consequentialism" (2:252). In response, he notes that a Telic Priority View will already have taken any significantly greater burden agents would have to bear into account. No doubt there will always be a way of "consequentializing" grounds for reasonable rejection and mirroring Contractualist Priority with a companion Telic Priority view. But even so, it will remain true that, according to Contractualism, only Contractualist Priority will take these burdens into account in the right way, namely, within the space of grounded claims.

[20] Parfit notes that Kant himself sometimes does say that rationality is a form of good (1:242).

wide value-based objective view of practical reasons will be able to proceed from Kantian Contractualism to Rule Consequentialism.

Naturalism. Space permits only a few observations about Parfit's careful and sophisticated engagement with forms of Nonanalytical (Metaethical) Naturalism. Begin, first, with a distinction all parties can agree to between normative-making properties and normative properties themselves. It is a familiar thought that if, for example, an action is morally wrong, the action must have properties that make it wrong: that it leads to bad consequences, that it is an instance of lying, or whatever. We can all agree also that, in the normal case, anyway, normative-making properties are natural properties. What is at issue between naturalists and their critics is whether normative properties themselves are natural. Now some versions of naturalism take the normative-making relation to be something like constitution. For such theories, if normative-making properties are natural, then so must be the normative properties, since they are made up of the normative-making natural properties. For example, suppose that hedonism is the correct theory of desirability or good. On most plausible views, this will be a necessary truth, so on this supposition the property of being a pleasure will be necessarily coextensive with being something there is normative reason to desire and so good in that sense. On some versions of Nonanalytical Naturalism, these necessarily coextensive properties would then be identical; the normative property of being a pleasure would be identical with being something there is normative reason to desire.

I agree with Parfit that these versions of Nonanalytical Naturalism are implausible. Although it is uncontroversial that anything with normative properties must have normative-making properties on which the normative property "supervenes," this is a conceptual rather than a metaphysical truth. Moreover, the normative-making relation is itself normative rather than metaphysical. So I agree with Parfit that any plausible Nonanalytical Naturalism must hold that anything with normative-making properties must also have a *different* normative property and that this normative property is itself also natural (2:295–305).

Let us focus now on Nonanalytical Naturalist views about normative reasons, taking as our example Subjectivist (existence internalist) views that identify practical reasons with what the agent would take as a reason and be motivated by under procedurally ideal deliberation. Internalist Naturalists must accept that the procedural ideal can be identified with the obtaining, in the limit, of such natural facts as being informed, dispassionate, (for some purposes) disinterested, and so on.

We should observe, by the way, that, as noted above, Parfit's own Convergence Claim also holds *reasonable belief* that normative facts exist, though not their existence itself, hostage to there being convergence on normative belief under like circumstances.

Nonanalytical Naturalists, as Parfit understands them, hold that normative facts and properties, though not normative claims, are, or can be reduced to, natural ones. In Chapter 26, Parfit argues plausibly against any Nonanalytical view that normative facts are natural on the grounds that the best way of identifying facts is by their information or content. If we think that a fact is *interesting* or *significant,* we must be identifying it by its content and not just referentially (2:336–41). But if that is right, then the Nonanalytical Naturalist, who thinks that normative claims cannot be reduced to naturalistic ones, should also believe that normative facts cannot be either.

Where Nonanalytical Naturalists usually make their stand, however, is with the claim that normative *properties* can be reduced to or are natural properties. Why can't they cheerfully concede Parfit's point about normative facts and hold the line on properties? Here it just seems much more plausible that the relevant properties can be identified referentially. None of this, of course, constitutes an adequate defense of Naturalism in the face of Parfit's careful and substantial critique, but perhaps it indicates lines along which such a defense might hope to proceed.

It is hard to imagine what could bring Nonnaturalists and Naturalists into agreement on fundamental matters of metaethics. (Though if Parfit's Convergence Claim bears out, it may be possible for both to get all they need by their respective metaethical lights to justify convergent reasonable beliefs about what matters.) When it comes to at least part of what matters, moral right and wrong, however, it seems that the possibility of agreement on what matters matters more. Although I have argued that Parfit's Convergence Argument fails, I nonetheless believe that his attempt to show that different traditions in moral philosophy (if not in ethics more generally) have been "climbing the same mountain" responds to a deep feature of morality's deontic dimension, namely, its consisting of standards to which we justifiably hold ourselves and one another in common.

References

Darwall, Stephen. 1983. *Impartial Reason.* Ithaca, NY: Cornell University Press.

———. 1990. "Autonomist Internalism and the Justification of Morals." *Noûs* 24: 257–68.

———. 1992. "Internalism and Agency." *Philosophical Perspectives* 6: 155–74.

———. 1997. "Reasons, Motives, and the Demands of Morality." In *Moral Discourse and Practice,* ed. Stephen Darwall, Allan Gibbard, and Peter Railton, 305–12. Oxford: Oxford University Press.

———. 2001. "Because I Want It." *Social Philosophy and Policy* 18: 129–53.

———. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge, MA: Harvard University Press.

———. 2010. "But It Would Be Wrong." *Social Philosophy and Policy* 27: 135–57.

Johnson, Robert. 1999. "Internal Reasons and the Conditional Fallacy." *Philosophical Quarterly* 49: 53–72.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

Pettit, Philip, and Michael Smith. 1990. "Backgrounding Desire." *Philosophical Review* 99: 565–92.

Railton, Peter. 1986. "Moral Realism." *Philosophical Review* 95: 163–207.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Harvard University Press.

———. 1980. "Kantian Constructivism in Moral Theory." *Journal of Philosophy* 77: 515–72.

Scanlon, T. M. 1998. *What We Owe to Each Other.* Cambridge, MA: Harvard University Press.

Schroeder, Mark. 2010. *The Slaves of the Passions.* Oxford: Oxford University Press.

Sidgwick, Henry. 1967. *The Methods of Ethics.* London: MacMillan.

Smart, J. J. C., and Bernard Williams. 1973. *Utilitarianism: For and Against.* Cambridge: Cambridge University Press.

Smith, Michael. 1994. *The Moral Problem.* Oxford: Blackwell.

Strawson, P. F. 1968. "Freedom and Resentment." In *Studies in the Philosophy of Thought and Action,* 71–96. London: Oxford University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127: 109–66.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments.* Cambridge, MA: Harvard University Press.

Watson, Gary. 1987. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology,* ed. F. D. Schoeman, 256–86. Cambridge: Cambridge University Press.

Robert Pasnau, *Metaphysical Themes, 1274–1671.*
New York: Oxford University Press, 2011. xi + 796 pp.

Some years ago Robert Pasnau (2002) began a book review in this journal by noting that "perhaps the most lively area of historical research in philosophy today concerns the scholastic antecedents of modern philosophy" (308). He went on to say, however, that "inasmuch as these [modern] authors were notoriously and proudly ignorant of scholastic thought, it is not to be expected that a better understanding of medieval and Renaissance philosophy will unlock the hidden meaning of modern texts." Now, with a magnificent book in which he examines a range of topics related to the metaphysics of substance, Pasnau has made research into the "scholastic antecedents of modern philosophy" livelier yet. He seems, however, to have changed his mind about the early modern philosophers' knowledge of their forebears; in the present work he repeatedly says that they got their scholastic forebears "largely right" (12) and "tend to know this material better than anyone does today" (72). He also now thinks that knowledge of late scholastic thought can shed light on the meaning of modern texts, a claim that is supported as persuasively by some of the discussions in this book as by anything I have seen.

Metaphysical Themes is large and covers much, so I cannot hope to touch on everything of interest. I hope to say enough to whet readers' appetites and indicate what one can expect to find in the book. It begins with the first critiques of the "classical scholasticism" of Aquinas and Bonaventure (the title's "1274" marks the death of both figures) and ends with what Pasnau sees as the first stage of postscholastic philosophy ("1671" marks the first draft of Locke's *Essay concerning Human Understanding*). Pasnau deliberately chose four centuries that cut across two historical periods in our usual narratives. Part of his ambit is to show that there is much in medieval philosophy that departs dramatically from stereotypical scholasticism and also that there are many remnants of medieval scholasticism to be found in such seventeenth-century figures as Descartes and Locke. Pasnau aspires to reform our usage of the term 'modern', reserving it for the period beginning circa 1900 and using nonprejudicial terms such as 'seventeenth century' for the historical period under discussion. I can only hope that his "perhaps quixotic" effort succeeds.

Pasnau does, however, think that the two most important philosophical trends in his four-century period, as a matter of both historical influence and philosophical interest, are scholastic Aristotelianism and mechanical philosophy. One might be tempted to move from seeing these as the two most important trends to thinking that there is a significant break between the medieval and early modern periods that justifies the standard periodization. Pasnau is right, however, not to wed doctrine to time period too closely since doing so obscures both medieval figures who were sympathetic to elements of mechanical philosophy and who formulated views that set the stage for its wider acceptance and also seventeenth-century (and later) figures who continued to defend Aristotelian accounts. Pasnau is, moreover, surely right to identify these two trends as especially important, particularly if our focus is on the metaphysics of material substance.

It is worth emphasizing that Pasnau treats all four centuries and both philosophical trends with care. Authors tend to focus on scholastic thought with some gestures to appearances of the claims under discussion in later postscholastic figures or to look at the scholastic figures purely as propaedeutic to the later figures they are really interested in. Not so in this book. Pasnau pays close attention to figures from throughout the four centuries, always analyzing arguments and positions as interesting in their own right. The philosophers who receive the most sustained attention are, in chronological order, Scotus, Ockham, Nicholas of Autrecourt (notable as an atomist in the heyday of Aristotelianism), Buridan, Nicole Oresme, Suárez, Hobbes, Gassendi, Descartes, Boyle, and Locke. Clearly, this is a book of interest to both medievalists and "modernists." I should add, too, that while that list of figures might be thought impressive enough, dozens of less well-known figures also play roles in Pasnau's story. If one mark of a good book in the history of philosophy is creating in readers a desire to read texts previously unknown to them, *Metaphysical Themes* succeeds brilliantly on that score.

As mentioned, Pasnau's focus is the metaphysics of material substance, a theme that opens more doors than one might initially expect. The book's thirty chapters are divided into six parts. Part 1 starts with the thought that all change requires a substratum that endures through change, and it continues with a discussion of the different theories of prime matter, which range from Aquinas's view that prime matter is nonextended and partless to Ockham's view that it is intrinsically extended. In this part, Pasnau also introduces atomism, along with an argument for the historiographical thesis that the history of philosophy consists not of a series of choices between Plato and Aristotle, but rather a series of choices between Plato, Aristotle, and Democritus, with Aristotelianism representing the middle ground between Plato and Democritus.

After a discussion of matter, one might expect a discussion of its Aristotelian complement, substantial form. Pasnau instead turns to substance in part 2 (though substantial form inevitably makes appearances), emphasizing the importance of distinguishing between substances in the thin sense (the per se unities composed of prime matter and substantial form and nothing else) and substances in the thick sense (thin substances plus a complement of accidents inhering in the former). Add the view that it is only accidents that we are directly acquainted with and it is easy to see why scholastic philosophers thought we could know little or nothing about the subject veiled by those accidents (or about the prime matter and substantial form composing that subject). Although corpuscularian philosophers rejected the hylomorphic analysis of substance, Pasnau argues that the scholastic doctrine of a veiled subject was remarkably persistent through seventeenth-century thought. Pasnau ends this part with detailed examinations of Descartes and Locke on substance. Perhaps most noteworthy is his spirited opposition to the common reading of Locke as positing a mysterious entity lying underneath ordinary substances. Pasnau argues that familiarity with the scholastic background makes it evident that Locke says nothing so strange. Instead, Locke introduces the term 'substratum' as a synonym for the traditional term 'substance', leaving him with a view that is banal rather than baffling.

With substances introduced, Pasnau turns his attention in part 3 to their properties or accidents, that is, the features that come and go while the substances persist. His chapters on inherence and modes are especially valuable, introducing a great deal of material that has been almost entirely neglected by scholars so far. Parts 4 and 5 are devoted to the "two principal kinds . . . of such properties": namely, quantity and quality. The topics covered in the part on qualities are reasonably standard ones that one might expect to see—real qualities and their significance, primary and secondary qualities, and powers and dispositions—but the part on quantity includes topics that one might not expect: the distinction between material and immaterial substance, location, and *entia successiva* (roughly, entities that cannot exist wholly at once but rather have one part at one time and another part at another). Here again Pasnau shines in giving systematic treatment to topics that are, as he rightly puts it, "woefully neglected." Nonetheless, this is one of the places where one could legitimately complain that Pasnau ought to have covered more topics than he did. He identifies quantity and quality as the two principal categories of the nine Aristotelian categories of accidents. But medieval philosophers usually understood that there is a third category worthy of special attention: relations. Relations seem especially relevant to Pasnau's project because in twentieth-century philosophy it is sometimes alleged that an Aristotelian account of substance closes the door on a satisfying account of relations, on grounds that at least some relations must be irreducibly polyadic but Aristotelian accidents can only be be monadic.

In part 6, Pasnau turns to issues of unity and identity, both synchronic and diachronic. It is here that he finally has a proper discussion of substantial form. As he notes, the scholastics can appeal to substantial form in trying to account for a substance's unity and identity through time. But when seventeenth-century corpuscularians deny substantial forms, at least for nonhuman material objects, they are forced to grapple with these issues without recourse to substantial forms and often end by giving up on commonsensical claims about natural kinds, the unity of substances such as trees and dogs, and diachronic identity.

One of the greatest virtues of *Metaphysical Themes* is the synoptic vision it presents of a period of the history of philosophy desperately in need of synoptic treatment. An occupational hazard of writing such works, of course, is becoming prey to the sniping of specialists who find this or that patch of the picture inadequate. I will indulge in two such shots, one more philosophical and one more historical.

Pasnau appeals in several contexts to the idea that Aristotelian hylomorphism is susceptible to both a metaphysical reading and a physical reading and that the later scholastics emphasized the physical reading, thereby setting themselves up for a fall when corpuscularians came along with a competing explanation for forms' reputed physical effects. In the case of substantial forms, Pasnau suggests that they can be conceived of metaphysically, as "abstract entities" that individuate substances as things of a given kind, or physically, as concrete entities that play a causal role in producing and sustaining substances' accidents and powers. A few passages suggest that Aristotle held the metaphysical conception, but Pasnau's considered view is that Aristotle bequeathed both conceptions to his followers. On Pasnau's telling, earlier scholastics emphasized the metaphysical role of forms, but later scholastics increasingly turned to the physical role. With the metaphysical role neglected, substantial forms were dismissed as otiose when the corpuscularians offered reductive mechanical explanations for substances' accidents and powers. Had the later scholastics emphasized substantial forms' metaphysical role, the mechanical philosophy might have seemed less promising.

It is no doubt true that scholastic philosophers appealed to substantial forms in a variety of contexts and that there are questions to be asked about how the different putative roles relate to each other. I am, however, not persuaded by Pasnau's attempt to sort these roles into two categories, metaphysical and physical. The contrast between forms as abstract, noncausal entities and as concrete, causal entities sounds stark, but on closer inspection it is difficult to see just what the division is or how it is supposed to work in the case of substantial forms. Pasnau wants to allow for thinking of substantial forms under both conceptions at once, but then the terminology is unfortunate, since little sense can be made of the same entity being both abstract and concrete. And why think that forms metaphysically conceived are abstract? It is hardly obvious that a principle of individuation, for example, would be on a par with numbers and propositions. I fare no better with the causal versus noncausal contrast. Clearly, 'causal' is not being used in the traditional Aristotelian sense since then no one would ever have doubted that substantial forms, aka formal causes, played a causal role. As for the modern sense of 'cause', I am not sure there is any single modern sense. One common view has it that causes are events regularly followed by events of a second type, but this seems an unpromising account for present purposes since forms are presumably not events in either conception. At one point, Pasnau suggests thinking of physical substantial forms as "internal efficient causes" (560). Given that scholastics commonly classified efficient causes as a species of external causes and formal causes as a species of internal causes, this description increases my worries rather than assuaging them. I am not merely quibbling about the term 'causes'. My suspicion is that once a sharply focused conception of causation is on the table, either some of the supposedly metaphysical roles will turn out to be causal or some of the supposedly physical roles will turn out to be noncausal.

Even if we leave aside the characterizations of the division, it is not clear to me that the roles on one side can or should be separated from those on the other side. It would be strange, for example, if there were not an intimate connection between something's being an enduring unity of a certain kind and that thing's having the accidents and powers that it has. The connection is readily explicable if a substantial form explains both. But suppose the former is explained by a metaphysical entity and the latter by a physical entity. Could God have combined the substantial form that explains my identity with whatever physical entity it is that gave my African violet its accidents and capacities? If that suggestion seems absurd, then maybe scholastic philosophers would reasonably object to Pasnau's attempt to divide "metaphysical" roles from "physical" roles. Furthermore, on the usual scholastic view, our only cognitive access to substantial form is through the qualities that, like other accidents, result from the form. If the qualities were not rooted in the substantial form in that way, then it would be even more difficult than it already is to see how we could know anything at all about substantial forms (cf. 634 and 648).

Moving on to my second worry, in the opening paragraph of the book Pasnau identifies one of the aims of his project as understanding the reasons for scholasticism's abrupt demise in the seventeenth century. That there was such a demise is part of our discipline's common lore, but is it true? Pasnau notes at one point that the "human mind tends to suppose that what it does not know about does not exist" (3). He has in mind the largely unknown philosophers of the fifteenth and sixteenth centuries, some of whom he is so admirably bringing to our attention. But I might suggest that there is a thriving scholastic tradition that goes well past the philosophers he labels "postscholastic." He cites the great number of scholastic texts from the centuries he identifies as terra incognita, but scholars who have taken the trouble to look could also cite a great number of scholastic texts from the late seventeenth and eighteenth centuries.[1] Yet scholastic authors later than Suárez are utterly absent from Pasnau's story. It is dif-

[1] Three scholars who have bothered to look are Knebel (2000), Novotný (2009), and Schmutz (2012). Pasnau's generally comprehensive bibliography does not cite these or any other works by these authors.

ficult to tell how philosophically rewarding these later scholastic texts are if no one reads them, but the mere fact that they exist makes scholasticism's seventeenth-century demise look rather mythical. It also leaves Pasnau's suggestion that Descartes saved philosophy sounding far-fetched.

But my criticisms are not meant to detract from the book's many virtues. Some books are worth reading; some books are even worth criticizing. Pasnau has written a book that belongs to the latter, more select group. I expect *Metaphysical Themes* to be a standard fixture on reading lists for many years to come and to spark many a journal article. And rightly so.

References

Knebel, Sven K. 2000. *Wille, Würfel und Wahrscheinlichkeit: Das System der moralischen Notwendigkeit in der Jesuiten-scholastik 1550–1700.* Hamburg: Felix Meiner.

Novotný, Daniel D. 2009. "In Defense of Baroque Scholasticism." *Studia Neoaristotelica* 6: 209–33.

Pasnau, Robert. 2002. "Review of Dennis Des Chene, *Life's Form: Late Aristotelian Conceptions of the Soul.*" *Philosophical Review* 111: 308–10.

Schmutz, Jacob. 2012. "Medieval Philosophy after the Middle Ages." In *The Oxford Handbook of Medieval Philosophy*, ed. John Marenbon, 245–66. Oxford: Oxford University Press.

Sydney Penner
Merton College, Oxford

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366521

Anita L. Allen, *Unpopular Privacy: What Must We Hide?*
New York: Oxford University Press, 2011. xv + 259 pp.

Anita Allen's recent book, *Unpopular Privacy: What Must We Hide?*, centers on a neglected area of privacy scholarship. Allen argues that there are areas of privacy that are fundamental and should be protected by liberal egalitarian governments despite the wishes of individual citizens. For example, the "don't ask, don't tell" policy adopted by the US military in the 1990s forced privacy on soldiers who may have wanted to disclose their sexual preferences. Within a liberal, feminist, and egalitarian framework, which is hostile to government interference with individual choice, Allen advances a moderate paternalism with respect to privacy. Allen writes, "We should live some of our lives in private, some in public; and there is often a role for government in requiring us to live this way. Privacy is too important to be left entirely to chance and taste" (196).

In the opening chapter, "Privacies Not Wanted," Allen notes that governments can justifiably mandate physical privacy by prohibiting neighbors from peering into our bedrooms. Similarly, governments mandate informational privacy by requiring secure health records, protecting professional confidences, and limiting the information sharing and gathering practices of online vendors. Some privacies are so important to individuals that they should be understood as inalienable. We don't allow individuals to sell themselves into slavery or sell their kidneys because these sorts of choices fundamentally undermine the dignity and freedom of these individuals. Allen understands privacy, along with personal freedom and race and gender equality, as a foundational political good that is necessary for individual well-being, dignity, and a just society.

The chapters that follow present a wide range of interesting cases and analyses—here are a few of Allen's conclusions. With respect to telemarketers intruding into the sanctuary of our homes, Allen endorses a ban on such practices (37). Banning such telemarketing is justified because some individuals don't realize the importance of privacy. Moreover, this sort of intrusion interferes with essential freedoms.

Concealment prohibitions on Muslim apparel such as the burka found in France are rejected by Allen as being unjustifiably paternalistic and discriminatory. While there may be times when the required removal of these coverings is justified, including while giving courtroom testimony and obtaining driver's licenses, Allen would make these cases exceptions and not the rule. Assimilation into the larger culture would not justify such practices. In this case, we should respect the religious choice of Muslim women to cover up.

On the other hand, Allen views restrictions on nude dancing as overly paternalistic unless such practices are also harmful and dehumanizing. If such actions degrade either the dancer or the patron, or lead to secondary effects like crime and drug use, then privacy should be mandated. Touching seems to be an important condition for Allen because "unless the woman is touched or confined, she cannot be overpowered" (91). Situations in which a dancer could be overpowered or physically controlled by a patron are demeaning and therefore should be subject to privacy regulations. Allen writes, "The rule against physical contact protects women from one particularly cruel, subordinating, dehumanizing danger, physical rape" (91).

Turning to informational privacy, Allen first considers confidential professional privacies. Part of maintaining a flourishing and free society includes guaranteeing the privacy of information that one shares with lawyers, doctors, and similar professionals. Allen notes that laws protecting medical information, for example, mandate privacy.

Racial privacy is also considered. Despite the fact that required disclosure of this sensitive information may have worrisome repercussions, Allen argues that racial privacy is outweighed by the legitimate government function of redressing past wrongs. But we have to be careful. Used to promote equality of opportunity or to redress past wrongs, such information is needed and useful. Nevertheless, Allen is aware that racial information can also be used to discriminate, oppress, and control.

She also claims that individuals freely give away too much information—we should be more aware of the value of privacy and personal information. Allen worries about the case of lifelogs and proposed "Total Recall" systems, or technology that may allow individuals to record every detail of their lives. Such systems could be used by corporations and governments to monitor or control populations. If privacy is morally valuable, then we should pause before adopting this technology without conditions. Among other protections, Allen recommends that no one should be required to keep a lifelog, that they should be considered the property of those who create them, and that owners of lifelogs should be able to delete content.

Allen concludes by defending the Children's Online Privacy Protection Act (COPPA 2000), which protects the privacy of children under the age of thirteen. For example, COPPA gives parents a veto over the "further use" of information collected from a child, but it also requires the security and confidentiality of this information independent of the wishes of anyone involved. Independent of the wishes of children, their parents, or internet operators, the act mandates privacy.

While there is much to praise in this book, I have two primary concerns. The first centers on Allen's conception of privacy, while the second deals with her justification of mandated privacy.

There are many privacies that Allen calls unwanted, unpopular, and coerced that I would call "isolation," or simply protecting rights and contracts. Consider privacy as isolation. One form of "mandated privacy" that Allen considers is house arrest or isolation in prison. If we view criminal activity as waiving one's liberty and privacy rights, then house arrest or isolation in prison would be chosen, not mandated. In good Kantian terms, we are using the principle that the criminal himself or herself has picked. Moreover, if privacy is valuable, then it would not include physiological and psychological forms of coerced isolation. If we include these forms of isolation in our conception of privacy, as Allen would, then privacy is not valuable full stop. Some forms of privacy are very harmful, and few would deny that adults or children who choose to endure these harmful privacies may be forced or nudged to take a different path, at least in some cases.

Or consider doctor and patient confidentiality. One could view this as a case where an individual's informational privacy rights over his or her own health information coupled with a general right to make contracts is protected by the law. Patients could broadcast their medical records on the evening news, thus waiving their privacy rights. Privacy is not paternalistically mandated in this case. This is also true of the confidences kept by lawyers, psychologists, and bankers.

When Allen does consider paternalistically mandated privacy—bans on telemarketing; "no contact" rules related to nude dancing; don't ask, don't tell policies; or protecting children online—I am unsure of the analysis. Why exactly should I not be allowed to opt-in to telemarketing? What fundamental or essential value am I unwisely tossing aside? More importantly, why is this loss so compelling that it would justify the government overriding my considered wishes along with those of the telemarketers?

I am also unconvinced of Allen's defense of "no touching" rules related to nude dancing. Women in these clubs may be as safe from rape and assault as women in other sex professions such as pornography. Moreover, the view that providing pleasure to another human being for compensation is dehumanizing or degrading needs defense.

Allen's critique of laws that would prohibit wearing burkas is also troubling for someone who champions liberalism, equality, and feminism. My worry is not so much with the conclusion that Allen offers, but with her reasons for attacking such laws. According to Allen, "modesty ought to be a right for those who consider it a core religious value" (75). My question is: why are religious folk so privileged? What if I, an atheist, donned a burka or an antimonitoring suit? Moreover, suppose that upon being asked why I would wear such a suit, I proclaim that it is my right to privacy and as long as I am doing nothing illegal or there is no special reason for me to take off the suit, it is no one's business who I am.

While there is much that I would disagree with in modern feminist gender theory, I would agree that there is something deeply troubling with ideological and religious worldviews being shoved down the throats of the young, especially views that lead individuals within these systems to be controlled and oppressed. If reasons matter, then it would seem that religious reasons for covering up should be no weightier than my secular reasons.

A final worry is that Allen's thesis is overly narrow. She claims to be offering an account of mandated or coerced privacy that is (1) consistent with a liberal, feminist, egalitarian democracy, and (2) promotes dignity and autonomy. This focus leaves aside the arguments and views of those who are not egalitarians or feminists. It would also have been helpful if Allen had provided a more rigorous argument for why violating the wishes of competent adults actually enhances liberty and dignity rather than debases it.

Despite my worries, Anita Allen's *Unpopular Privacy* is a welcome addition to privacy scholarship. Her conclusion is surprising: some types of privacy are so important that we should prohibit children and adults from waiving their rights. For over twenty-five years, Anita Allen has written about privacy and at the same time influenced a host of scholars across numerous disciplines. In this volume, Allen offers numerous cases, philosophical arguments, and enlightening analysis. *Unpopular Privacy: What Must We Hide?* deserves and will benefit a wide readership.

Adam D. Moore
University of Washington

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366529

Marc Lange, *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature.* New York: Oxford University Press, 2009. xviii + 257 pp.

In recent years a resurgence of interest in the laws of nature has led to a number of new books on the topic, articulating a variety of perspectives from neo-empiricism to neo-Aristotelianism, and from primitivism to eliminativism. Of these new works, in my view, the most original and intriguing is Marc Lange's *Laws and Lawmakers.* It has long been noted that laws support counterfactual and subjunctive conditionals. Generally those who have thought this an important relationship have held, as would seem natural, that laws are prior to the conditionals. Nonetheless, Nelson Goodman pointed out the difficulty of straightforwardly deriving the conditionals from laws plus categorical facts. David Lewis's metaphysics has the conditionals fixed by the proximity structure of possible worlds, which is itself in turn fixed in part by the laws—similarity of laws trumps similarity of nonnomic facts in making worlds close. But it is not clear what metaphysical reason there is for supposing *regularities*, which is what Lewis's laws are, should have this trumping power. So, despite the fact that consensus gives long odds to the view that the conditionals are fundamental, the smart money might be on that option. Hence the prima facie ground for taking seriously Lange's proposal that subjunctive facts are primitive and that we can explain which facts are the laws in terms of the subjunctive and counterfactual conditionals (I will henceforth use 'subjunctives' for both).

We take laws as held fixed, as far as possible, when considering subjunctives. The consensus view holds that this is because the subjunctives are determined by the laws (and nonsubjunctive facts). But maybe it is just that the nomic facts are especially stable under subjunctive considerations. Some facts are more stable than others under hypothetical subjunctive circumstances. If the temperature in Bristol were to remain below 0°C for a couple of days, then the water in the pond would freeze. But it would still be colder in Calgary than in Bristol (this being written in January). If it were to be colder in Bristol than in Calgary, it would still be the case that water freezes at 0°C. Perhaps laws remain true under any subjunctive circumstance: L is a law if for any F, F $\Box\!\!\rightarrow$ L. That cannot be quite right, for the following looks to be true: if gravity were an inverse cube law, then the strength of gravitational attraction would fall off even faster with distance. But note that in this case, the antecedent is inconsistent with the actual laws. So perhaps the laws remain fixed under subjunctive conditions consistent with the laws, unlike the nonnomic facts. This is the essence of Lange's account of the laws of nature.

Consider all the facts not including those articulated using phrases such as 'it is a law that . . . ' (that is, including 'gravitational force is inversely proportional to the square of distance', but excluding 'it is a law that gravitational force is inversely proportional to the square of distance'). These are the 'subnomic facts' in Lange's terminology. A subset of this set is 'sub-nomically stable' if every subjunctive supposition consistent with that set leaves members of the set unchanged (as in: 'if the Earth were twice its current size, gravitational force would [still] be inversely proportional to the square of distance'). Lange proves (very nicely) that the subnomically stable sets form a hierarchy of nested sets. The set of all subnomic facts forms the largest (and trivially) subnomic set. The next subnomically stable set is one containing the laws, plus mathematical, conceptual, and logical truths—it excludes all the accidental truths. A set containing the mathematical, conceptual, and logical truths but not the laws is a subnomically stable subset of this set and so forth. This hierarchy is a hierarchy of degrees of necessity. Compared to the accidental truths, the laws of nature are necessary. But the mathematical truths are even more necessary because they are part of a set that is stable under counterlegal suppositions: even if gravitational force were inversely proportional to the cube of distance, two squared would still be four.

It is this hierarchy of sets of necessities that gives Lange's view an advantage over competing views, in particular the view that lawmakers are powers. That view makes the laws necessary also, but gives them full-on metaphysical necessity. That view does not allow for degrees of necessity—metaphysical necessity is the only real necessity, and either a fact is necessary or it is not. So the powers view squashes Lange's hierarchy, preserving only the distinction between the accidents and the laws. Lange regards it as an advantage that his view neatly accommodates the appearance that mathematical truths are more necessary than nomological truths. Furthermore, it allows for a distinction within the laws, for some laws seem to be more resilient under counterfactual suppositions than others: the symmetry and conservation laws seem stable under counterfactual changes to the force laws: if gravitational force were inversely proportional to the cube of distance, then energy would still have been conserved.

There is a great deal in this book. Lange discusses, for example, how his account handles (better than other views) the "lawmaker's regress" (a regress in accounting for the necessity of laws), the relationship between laws and objective chances, and the nature of instantaneous velocity and acceleration. The latter, argues Lange, are best understood in terms of certain subjunctive facts. This, in turn, is one advantage of what might otherwise seem to be a disadvantage, the fact that his ontology rests on a basis of subjunctive facts, with laws as derivative—whereas almost every other metaphysician wants to make subjunctives derivative. While Lange makes an excellent case for his preference, it does nonetheless leave us with a quantitative profligacy. There are *lots* of (very specific) subjunctive facts. If laws are basic, then we can (hope to) explain the subjunctive facts with a handful of fundamental laws. I need to be told more about how subjunctive facts organize themselves if I am to be comfortable with so many of them.

Lange writes with restrained elegance, but this does not disguise the fact that his book is in many places hard intellectual work. That reflects the fact that this is a work of highest intellectual caliber, not unlike important advances in mathematics, that happens to deal with difficult material. It is to Lange's credit both that his cleverness produced such ideas and that he makes it as easy as he can for the reader to follow them. The reader's effort will be amply repaid. This book is a must for anyone interested in laws, but I would recommend it to any metaphysician also since Lange's view has ramifications well beyond the realm of laws and his book is exemplary of how we should approach our work.

Alexander Bird
University of Bristol

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366537

Terence Irwin, *The Development of Ethics: A Historical and Critical Study*. Vol. 1: *From Socrates to the Reformation.*
Oxford: Oxford University Press, 2007. xxvii + 812 pp.

This volume, the first of a series of three, is an outstanding example of high scholarship combined with deep insight in ethics as a philosophical discipline. Terence Irwin has published several books and papers on Plato and Aristotle that have become central to academic reflection on Ancient philosophy in the last three decades. This new work is a landmark in academic reflection on ethics, ranging from Socrates to Reformation (vol. 1), from Suárez to Rousseau (vol. 2), and from Kant to Rawls (vol. 3), thus embracing in a single undertaking the main ethical doctrines that Western philosophers produced over many centuries. This is such a formidable task that almost no one would think of carrying it out, but a task that in the current case pays off awfully well—or so I want to argue in what concerns the first volume, which is focused on Ancient and Medieval philosophy.

What looks fearsome is not exactly the huge number of books, texts, and doctrines one needs to go through. What is really frightening is the search for a common thread, *ein roter Faden* so to say, by means of which the whole discipline of ethics would be unified as a long but single conversation that can be evaluated in terms of a set of tenets being discussed and developed in all these thinkers. This was done for logic by Kneale and Kneale in the 1960s, and it is understandable that they succeeded in finding such a thread, for contemporary logic, the logic stemming from Frege's work, allows us to see and evaluate its history from Aristotle onward as steps to, or deviations from, what counts as a sound logic theory. One has thus a development of logic—but is there something similar in ethics? One may reasonably doubt this and think instead of ethics as a discipline full of controversies, perhaps even as being forever controversial, as no resting point is ever to be found in it, so long as (moral) values are in a relevant way subjective or historical, quite the opposite of what happens in natural science, let alone in logic or mathematics. Now, Irwin has a clear answer to this worry. Irwin's *roter Faden* is what he vindicates as "Aristotelian naturalism":

> [Aristotle] defends an account of the human good as happiness (*eudaimonia*), consisting in the fulfillment of human nature, expressed in the various human virtues. His position is teleological, in so far as it seeks the basic guide for action in an ultimate end, eudaemonist, in so far as it identifies the ultimate end with happiness, and naturalist, in so far as it identifies virtue and happiness in a life that fulfils the nature and capacities of rational human nature. This is the position that I describe as 'Aristotelian naturalism', or 'traditional naturalism'. We can follow one significant thread through the history of moral philosophy by considering how far Aristotle is right, and what his successors think about his claims. (4)

One has thus *a* thread, and a very significant one; but is this *the* thread, that *roter Faden* by means of which we will be able to construe the entire history of ethics as the development of a basic moral position? As significant as it may be, it cannot work as the dividing line as in the case of the development of logic, for such a position will not be fair to many moral doctrines that react against Aristotle's naturalism on ethics—and sometimes react very strongly against it, as in the case of Hobbes. But the idea is precisely to invert the claim: with Aristotelian naturalism we have such a robust moral theory that the history of ethics may be seen as the unfolding of different reactions to it, either in developing it and building on it, or in rejecting it in many ways. Ethics becomes, so to say, a history of putting Aristotelian naturalism to the test, and somehow it must withstand the test—but this belongs to the third volume to show. What is crucial is that, even though there is no unique point from which we can adjudicate the past, there is a moral position that stays salient enough to make all other moral doctrines turn in some way around it. The history of the unfolding of these reactions may be seen as the development of an ethical position that gets stronger by eliminating weak parts, as well as by taking on accretions—or so is the bet.

To get there, three moves at least seem necessary. First, Aristotelian naturalism should be construed in a quite generous way. That is, commitments to teleology, for instance, must be weakened as far as possible, such that attacks on teleology—and the fatal blow thrust by Darwin—will not be lethal to it. Only teleology in action is to be preserved. But this is no desperate task, as Aristotle himself was eager to avert explicit dependence of his ethics on metaphysics or natural teleology. A second move is to strengthen parts of Aristotle's ethics. Some notions seem to be lacking in his ethics, such as the notion of will, or not clearly presented, as the notion of intention, and these notions play a crucial role in our discussions on moral responsibility. Aquinas is one of the privileged authors in this sense—actually, he gets the biggest part of this volume. And this is no coincidence, for Irwin takes Aquinas as offering the best reformulation of Aristotle's naturalism: "The best way to examine this [Aristotelian] approach and this naturalist position is to reflect on Aquinas' version of them. For this reason, my chapters on Aristotle omit some questions that one might expect to see discussed there; I postpone them until I discuss Aquinas and his critics" (4).

It may well be the case that Aquinas's reading of Aristotle's ethics gives (Aristotelian) naturalism its strongest version—taking for granted that Aquinas's own religious compromises have no direct effect on it. But a third move is still necessary: to smooth down those reactions that look rather as outright rejections of Aristotelian naturalism. Eudaimonism is central to Aristotelian naturalism; but Kant argued bitterly against it, as he took it that any eudaimonism is the *euthanasia* of all morality. There has been recently an attempt to bridge the gap between Kant and Aristotle, and there are good reasons to do so. Still, it demands a good deal of argumentation not to see them as conflicting ways of explaining the phenomenon of morality. Nietzsche and his genealogy of morals should also figure prominently in such a list.

Granted all these moves, the outcome of volume 1 is an extraordinary critical analysis of Ancient ethics, beginning with the predecessors of Aristotelian naturalism (such as Socrates and Plato) and going through its legacy, as well as with those doctrines that found in Socrates or Plato reasons to resist some Aristotelian tenets, or those that brought about new topics. Irwin displays a superb knowledge of sources, conducting the discussion in chapters according to a historical line, but always with an eye to the conceptual formulation and evolution of ethical notions. There is so much scholarship, subtlety, and ingenuity in Irwin's analyses that I can only mention them here. I will focus on three points that seemed to me central to Irwin's strategy of positing Aristotelian naturalism (as seen by Aquinas) right in the center of moral reflection. By doing so, I hope to give the reader a slight idea of the richness of this book.

The first point concerns the Aristotelian notion of *prohairesis* and its rival Stoic notion of *sunkatathesis.* According to the Stoics, after apprehending an object by *phantasia,* the agent gives or refuses his or her assent, and from this assent an impulse is generated leading to action. Now, human assent is based on reasoning, and this reasoning may contain deliberations about how to act. Stoics do not seem to have put much emphasis on deliberation, but deliberation is expected to occur whenever the agent has to give his or her assent to an impression and look to put the resulting impulse into practice. Deliberation is at the center of Aristotle's reflection on human action. Deliberation concerns what is up to us to do or not to do, and one may construe this notion basically in two different ways: either in a libertarian way, as Alexander of Aphrodisias did, in the sense that what is up to us to do now is also up to us not to perform; or in a compatibilist way, in the sense that it is determined now, given the circumstances and the desires we have, that we will do what we are about to do, albeit, if the circumstances are altered, or our desires, we can act otherwise. In this latter sense, the agent has a general capacity of acting otherwise, whereas in the former sense, the agent has the specific capacity of acting otherwise in these precise circumstances, *hic et nunc,* as he is about to act. The libertarian reading is incompatible with the Stoic notion of *fatum* and makes Stoic assent a rival to Aristotelian *prohairesis,* thus proposing a competing conception of moral responsibility. Since Richard Loening's *Die Zurechnungslehre des Aristoteles* (1903), and notably after Susanne Bobzien's *Determinism and Freedom in Stoic Philosophy* (1998), there is a strong tendency among commentators to saddle Aristotle with a compatibilist version of what is up to us. This is the way Irwin reads this notion:

> Aristotle's account of responsibility is closer to the Stoic position than to Alexander's. While he affirms the conditions that Alexander interprets in an indeterminist sense, he does not interpret them in this sense. He argues that rational agents are rightly held responsible for their voluntary actions because these are actions of agents who are capable of rational deliberation and election. The Stoics go further, and insist that we are responsible for those actions that actually express—either by reflexion on appearances or by simply going along with them—the outlook that is embodied in the agent's elections (as Aristotle understands them). This further element in the Stoic position modifies and develops Aristotle, but it does not depart sharply from his position. Later expositors of Aristotle's position are right, therefore, to mention assent. (308)

This reading may well be right and has recently been favored by important commentators (see, for instance, Michael Frede's *A Free Will* [2011]). I don't want to discuss it as such; my point is that such a reading gives support to, but is also supported by, the main idea of a sort of continuity throughout ethical thought. Alexander's libertarianism is thus put aside as a divergent interpretation of Aristotle's thesis—but there goes with him the Principle of Alternate Possibilities as central to moral responsibility—and this is a crucial move.

A similar attitude is taken concerning the connected idea of *will*. This notion has not been introduced by Augustine in a sheer break with Ancient thought; quite the contrary, Augustine "does not commit himself to any claims about the will that are inconsistent with the Stoic view of assent" (411). Augustine is rather on the track of intellectualism: "The will is free in relation to the passions in so far as it is capable of consenting or not consenting to the actions suggested by the passions. The will is not similarly free in relation to the apparent good, but Augustine does not suggest that this lack of freedom involves any lack of the freedom relevant to responsibility" (412). Or, as Irwin said some lines before: "We have no reason, therefore, to attribute voluntarism to Augustine. He emphasizes the role of the will in free and responsible action because he believes that non-rational desires can move us to free action only with our consent; he does not claim that the will moves us independently of the greater apparent good. He accepts Stoic intellectualism and avoids voluntarism" (412).

Again, there is continuity; and this may well be right. But continuity may look a bit Procrustean. Let me bring in my second point to elaborate this idea. Aristotelian *prohairesis* is chiefly concerned with the means to an end, or so insists Aristotle. But we think rather that it should be mainly concerned with establishing the ends we pursue, and only subsequently with the means to reach these ends. There are different strategies to cope with the restriction of deliberation to means in Aristotle. Aquinas has convincingly argued that the final ends (or pleasure, honor, and knowledge, according to the Aristotelian tripartition) may be taken as means to (in the sense of components of) happiness. In this way, one may have different ends *materialiter* speaking, but only one *formaliter*, namely, happiness. This point is highlighted by Irwin in the following way: "Rational agents accept an ordered plurality of ends, and want the satisfaction of their desires to correspond to the comparative value they attach to each end. . . . To adjust one end to others is to recognize the structure of an ultimate end embracing them all; hence the rational pursuit of any particular end for its own sake requires its subordination to an ultimate end" (496).

Aquinas pursues further the idea of practical wisdom not only in the subordination of our ends to the ultimate end, but also in determining those ends that are final to our actions. These will be the first practical principles, in relation to which practical wisdom is also expected to fix the right means to achieve them. Irwin claims that Aquinas provides the best reformulation of Aristotle's ethical thought; we have thus to reconcile this double role of practical wisdom with the Aristotelian doctrine of deliberation as the central practical usage of reason. As there are passages in Aristotle that point in this direction, continuity seems to be on the right track. But continuity here means bringing in the notion of *synderesis*, and *synderesis* is what corresponds, in practical reason, to the noetic apprehension of first principles in knowledge. So we now have apprehension of ethical first principles, and these principles are universal and apprehended by all of us. This gives the ultimate end much more definite content than one would suspect it may have in Aristotle; moreover, deliberation, which was the central aspect of practical reasoning in Aristotle, has no significant role to play in the apprehension of these practical principles. The universal conscience, claims Aquinas, is indestructible, for it defines the human agency in itself. That goodness is to be pursued and evil to be averted is a mark of human rationality and not a psychological or anthropological outcome, as Irwin highlights. There are huge consequences here. Malignity is now out of human reach, as Kant will claim later; and we as rational agents can do evil only as deviation from the moral rule, which at the same time we recognize. Cruelty becomes a sort of stepping aside from moral law and not an object of choice by itself. Moreover, we are heading to universal laws in morality: "This division between two roles of practical reason has no explicit Aristotelian support, but it is a reasonable expansion of Aristotle. For we need some account of how we can form the ends that are characteristic of the virtuous person. To answer this question, Aquinas introduces synderesis ('universal conscience'), which is the specific disposition of practical reason that grasps the first principles. He claims that these principles are the first principles of natural law" (573).

Is Aristotelian naturalism still there, with universal natural laws (Aristotle thought of natural justice as a case of political justice), or has it been so altered by these accretions that it is barely recognizable, somewhat like the sea-god Glaucus in *Republic* 10? I come to my third and last point. As Aristotelian eudaimonism is self-centered, one may ask how to incorporate into it the notion of altruism, which seems to be so central to moral thought. Now, Aristotle's doctrine of friendship does have such a notion, for a friend looks after other people for the sake of themselves, and not because his or her doing so may be good, useful, or pleasing to himself or herself. This is altruism, but limited in that that these other people are restrictedly his or her friends. Worse, your friend is, according to Aristotle, another yourself, albeit separate: altruism is severely restricted to replication of one's own self. Compare Aristotle's altruism to the New Testament's parable of the Good Samaritan: we seem to be poles apart. One may extend Aristotelian altruism toward all the members of the *polis*, but there still seems to be lacking a requirement of impersonality and detachment, which stands out so prominently in that parable. Irwin is well aware of this:

> The more we are inclined to associate morality with impartiality, impersonality, and detachment, the more surprised we will be by Aristotle's treatment of it. For while he takes seriously the requirements of justice and fairness, he tries to derive them from self-regarding and self-centered concerns; the close connexion between friendship, the com-

> mon good, and justice shows his preferred direction of argument and justification. This is not merely a theoretical difference from other ways of thinking about morality; it also affects the moral principles that Aristotle accepts and emphasizes. Duties are owed to other people as friends and fellow-citizens sharing goals and interests with the agent, not simply as other people. Non-members of a community have no clear moral claims on me. The human beings or nearly-human beings who cannot be fellow-members of a community are legitimately treated as natural slaves and used as instruments for my benefit rather than theirs. (231)

Let us say that Aristotelian altruism goes for a possible altruism whenever other people are somewhat like us, whereas the Good Samaritan imposes a necessary altruism, whoever happens to cross our way. This possible altruism seems hard to reconcile with Kant's universal human agent, or even with utilitarian rules of maximizing the benefits for everyone, for both of them respond to the generalized claim of (necessary) altruism. Is there a way to see here some continuity nonetheless, or do we face a break in ethical thinking, a new basic requirement being introduced that has no trace in Aristotle's thought? Here Irwin seems to me to make an excellent point. First, he shows how this idea goes back to the Stoic notion of *honestum* and how the Aristotelian notion of acting *tou kalou heneka* may be seen as the origin of it. But second, he shows also why we should be attentive to what looks at first glance as an unwarranted restriction of altruism in Aristotle: "This egocentric aspect of Aristotle's view does not necessarily indicate an error in it; indeed, it may be a theoretical advantage. For it explains why we might recognize a more stringent requirement corresponding to our closer connexion with some people than with others" (232).

There are now features of a moral action, such as detachment, impartiality, universality, that have to be coordinated, and may occasionally conflict, with others—like pursuing one's own interests—all under the heading of one's doing well, instead of taking some of them as more basic and explaining away the others in terms of them. Let me quote Irwin again:

> We probably cannot decide unequivocally that Aristotle's conception of morality does or does not match ours; for we probably lack any pre-theoretical conception of morality that is definite enough either to agree or to disagree with Aristotle, or with Bentham or Kant. Rather, our beliefs about morality include some that Aristotle may plausibly claim to explain, as well as others that do not fit his account. Aristotle's conception of morality is not inaccessibly remote from ours. His explanation of morality may advance our understanding of it. (232)

We are thus back to our initial issue: continuity, or breaking point? One may also ask: Is this expansion toward universality and impartiality an outcome somehow in agreement with ethical naturalism, or is it a requirement derived from a theological perspective, according to which we are all members of one big family, and thus hard to square with Aristotelian doctrine? Does moral discourse make for a coherent language, or is it a quilt made up from different threads? Irwin argues strongly for a coherent, evolving language that starts with Aristotelian naturalism and develops toward more sophisticated thinking as it incorporates notions that strengthen that original position. He argues brilliantly for this position, displaying an enviable scholarship, a sharp analytical account, and such a mastery of sources and texts that make this book a wonderful and indispensable reference for any work in ethics as a philosophical discipline.

Marco Zingano
University of São Paulo

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366545

Theodore Sider, *Writing the Book of the World.*
Oxford: Oxford University Press, 2011. xiv + 336 pp.

Writing the Book of the World is masterful work. Sider offers a careful and insightful treatment of metaphysical structure (or naturalness, or fundamentality), developing a framework for metaphysical inquiry while defending the centrality of metaphysics to philosophy as a whole. There is much to admire and much to discuss. I expect Sider's book to become a touchstone for further discussion of the nature and viability of metaphysical inquiry. It is required reading.

Sider's project is a radical extension of Lewis's (1983) project in "New Work for a Theory of Universals." Lewis—building on Armstrong's (1978) defense of universals—identifies theoretical roles for a metaphysical posit of *natural properties,* while pointing out that one need not posit universals to fill these roles: tropes will do, as will a primitive status of naturalness for elite sets of possibilia. Sider's (vii, 8, 85) core innovation is to extend the Lewisian status of naturalness "beyond the predicate," allowing it to apply to any portion of the language, including quantificational phrases and logical connectives. Thus Sider (92) posits a primitive "structural" operator, which maps arbitrary portions of the language to truth if and only if that portion of the language "carves at the joints" (or "is perfectly natural," or fits "the fundamental structure of reality"). Given that a fundamental theory of reality should be cast in all and only joint-

Thanks to Marco Dees, Janelle Derstine, Jeffrey King, Ted Sider, Tobias Wilsch, and Dean Zimmerman.

carving terms (106–9), what Sider gives us is a device for querying whether a given portion of the language should appear in "the book of the world."

Sider's own book divides roughly in thirds. In the first part (chapters 1–5), Sider identifies a wide range of roles for his extended notion of structure. In the second part (chapters 6–8), Sider develops a positive theory centered on his primitive "structural" operator. And in the third part (chapters 9–13), Sider considers applications of his theory to first-order questions about ontology, logic, time, and modality, culminating in a sketch of a radically eliminativist worldview on which, fundamentally speaking, there are only spacetime points and sets.

I will focus on one critical point, concerning a mismatch between the senses of 'structure' in play in the first and second parts of Sider's book:

> *Mismatch*: The roles for structure are for "structural enough" and "more structural than," but Sider's primitive is "perfectly structural."

Given *Mismatch,* Sider's primitive does not play the roles it was posited for. This would be bad news for Sider since he recommends structure as "a *posit*. . . justified by its ability to improve our theories of these matters" (10).

The roles for structure are for "structural enough" and "more structural than": This is implicit in Sider's early illustrations involving colors (1–2) and chemicals (6–7) and becomes explicit when Sider speaks of reference as "joint-carving" (28) and allows that there can be substantive disputes in the special sciences (48). Evidently there can also be laws, explanations, confirmation, and induction in the special sciences as well, and presumably there is greater epistemic value in chemistry than in some gruesome counterpart ("schmemistry"). Sider explicitly acknowledges this point in passages such as (141): "Genuineness of explanation does not require *perfectly* structural notions, as we see from the special sciences." Overall, the roles Sider invokes reflect a need to distinguish predicates like 'green' from gruesome counterparts like 'grue'. One wants to say that 'green' is more natural than 'grue', and natural enough for induction on the color of emeralds, but of course 'green' is not perfectly natural.

Sider's primitive is "perfectly structural": His primitive "structural" operator is explicitly described (128) as fitting an absolute notion of perfect structurality. His operator only maps terms from fundamental physics (and perhaps mathematics) to truth. From the perspective of the book of the world, 'green', 'grue', 'helium', and 'schelium' all go equally unmentioned. Indeed this restriction to the perfectly structural is a consequence of Sider's guiding requirement of "purity" (106), on which (to speak in parables): "When God created the world, she was not required to think in terms of nonfundamental notions like city, smile, or candy."

Thus *Sider's primitive does not actually play the roles it was introduced to play.* The roles for "structure" are roles for a natural/gruesome distinction, but Sider's primitive instead draws a fundamental/nonfundamental distinction.

The problem underlying *Mismatch* is that these distinctions do not match: the natural/gruesome distinction comes in degrees and still applies in nonfundamental domains like chemistry.[1]

All might still be well *if* Sider could use his primitive notion of "perfectly structural" to define the needed notions of "structural enough" and "more structural than." Then even though "perfectly structural" would not play the roles for structure, it would still yield an understanding of the notions that do play these roles. Sider is well aware of this. Indeed matters come to a head in the space of two paragraphs on page 129, beginning when Sider acknowledges: "[W]e need a comparative notion of structure in many of the applications." He adds: "Talk of comparative structure must have metaphysical truth-conditions in terms of absolute structure." But he immediately continues: "How to give such metaphysical truth-conditions? How to define comparative structure? I do not know." This is laudably candid. But what is it but an admission that, by Sider's own lights, it is unknown whether his primitive is of use?

Three main options seem open for the metaphysician who would follow Sider in invoking structure, in order to fix *Mismatch*. First, she might revise the roles for structure, so as to find roles for Sider's "perfectly structural." Perhaps this is a useful primitive for other reasons.[2] Though, by my lights, Sider's first five chapters make a compelling case for "structural enough" and "more structural than." This first option would still leave one without an account of these needed notions.

Second, the friend of structure might devise an indirect match, by defining "structural enough" and "more structural than" in terms of "perfectly structural." This is Sider's option—the one he admits not knowing how to achieve. And how to achieve this? If one starts with just a cut between the fundamental and the nonfundamental, how can one recover any of the "structure" within the nonfundamental? How can one go on to draw any distinctions between 'green' and 'grue', or 'helium' and 'schmelium', if all get lumped together as nonfundamental from the start?

Sider does offer some suggestions toward this second option (129), starting from an extension of Lewis's idea of length of definition and encompassing several further ideas. I am skeptical. But for present purposes suffice it to say that a series of suggestions is not enough even by Sider's own lights (117). By Sider's lights, one owes at minimum "toy metaphysical truth-conditions" for "structural enough" and "more structural than" that could "convince us that

[1] *Mismatch* is in fact a problem that Sider inherits from the Armstrong-Lewis framework he would extend.

[2] Of the dozen roles that Sider presents over his first five chapters, I see two that might plausibly be understood via "perfectly structural": intrinsicness (10) and spacetime structure (38–43). So perhaps "perfectly structural" is still useful for at least these two purposes.

there is a real metaphysical semantics, even if that metaphysical semantics is too complex for us to discover." It is a substantive claim that a given notion has a metaphysical semantics in certain specific terms.[3] This second option would leave one facing the task of substantiating this claim.

As a third and final option, the friend of structure might revise her stock of primitives, perhaps adding primitive operators associated with "structural enough" or "more structural than." Indeed it seems that *a primitive "more structural than" operator can go it alone.* A "more structural than" operator provides strictly more information than a "perfectly structural" operator. While the latter only induces a cut between the fundamental and the nonfundamental, the former (given plausible assumptions) induces a richer partial ordering structure, against which notions like *minimal element* and *rank* are definable. So one need only say that a notion is perfectly structural if and only if no notion is more structural than it (it is a minimal element in the ordering), and that a notion is structural enough if and only if it has a low enough rank (assuming a natural ranking function over the ordering). Then one will have recovered all of the relevant senses of 'structure'.

But Sider objects to this third option (129), primarily on grounds that it violates his purity constraint. Sider argues that the primitive must itself count as perfectly structural and should itself feature in the book of the world (138–41). But a primitive "more structural than" cannot feature in the book of the world, at least in the most straightforward way via a sentence of the form 'Φ is more natural than Ψ', without the purity of the book thereby becoming sullied by the imperfectly natural 'Ψ'.[4] So this third option would leave one needing to reconsider purity.

Putting this all together: The senses of 'structure' in the first and second parts of Sider's book are mismatched, and one who who would fix this mismatch has her work cut out for her.

Criticism aside, *Writing the Book of the World* is outstanding work, and required reading for anyone interested in the nature and viability of metaphysical inquiry. Sider's book sets the terms of the debate. Indeed, just as Sider's lauded *Four-Dimensionalism* (2001) helped set the agenda for the previous decade in metaphysics, I expect *Writing the Book of the World* to help set the agenda for the next decade.

[3] Indeed Sider notes that it is a substantive claim that "causation" has a metaphysical semantics in noncausal terms (117–18).

[4] Sider (2009) uses a comparative "more natural than" operator. The view in *Writing the Book of the World* thus constitutes a change in view for him, mainly driven by the demand for purity. (A second change in view: Sider's preferred mereology has shifted from the universalism of Sider 2001 to nihilism, driven by the demand for ideological parsimony, which invites the elimination of mereological terms if possible.)

References

Armstrong, David. 1978. *Universals and Scientific Realism.* 2 vols. Cambridge: Cambridge University Press.

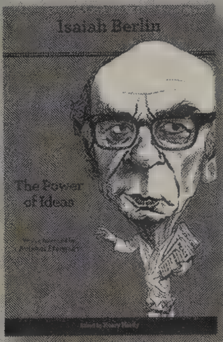
Lewis, David. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61: 343–77.

Sider, Theodore. 2001. *Four-Dimensionalism.* Oxford: Oxford University Press.

———. 2009. "Ontological Realism." In *Metametaphysics*, ed. David Chalmers, David Manley, and Ryan Wasserman, 384–423. Oxford: Oxford University Press.

Jonathan Schaffer
Rutgers University

Philosophical Review, Vol. 123, No. 1, 2014
DOI 10.1215/00318108-2366553



The Power of Ideas

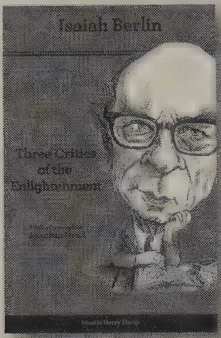
Second Edition

Isaiah Berlin

Edited by Henry Hardy

With a new foreword by Avishai Margalit

Paper $24.95 978-0-691-15760-3



Three Critics of the Enlightenment

Vico, Hamann, Herder

Second Edition

Isaiah Berlin

Edited by Henry Hardy

With a new foreword by Jonathan Israel

Paper $29.95 978-0-691-15765-8

Concepts and Categories

Philosophical Essays

Second Edition

Isaiah Berlin

Edited by Henry Hardy

With a new foreword by Alasdair MacIntyre

Introduction by Bernard Williams

Paper $24.95 978-0-691-15749-8

Karl Marx

Fifth Edition

Isaiah Berlin

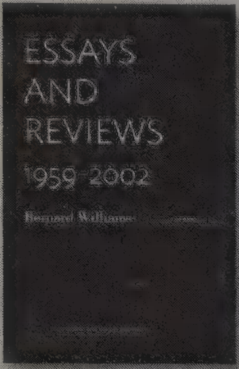
Edited by Henry Hardy

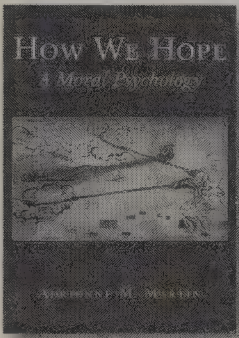
With a foreword by Alan Ryan and an afterword and guide to further reading by Terrell Carver

Paper $24.95 978-0-691-15650-7

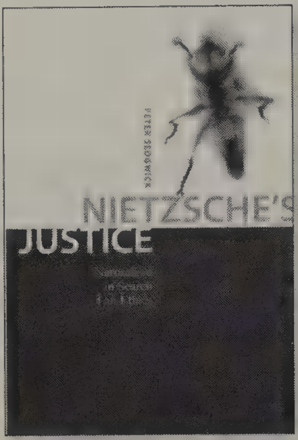
PRINCETON UNIVERSITY PRESS

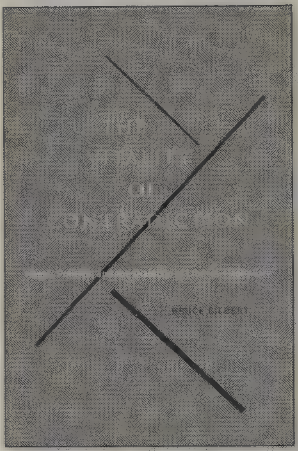
See our E-Books at
press.princeton.edu


ESSAYS
AND
REVIEWS
1959–2002
Bernard Williams

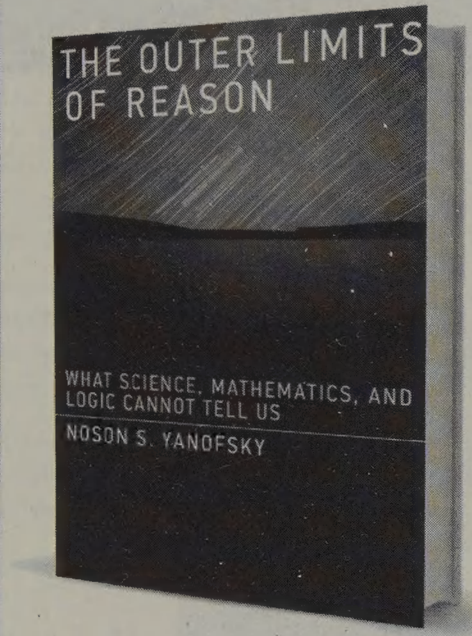

HOW WE HOPE
A Moral Psychology
ADRIENNE M. MARTIN

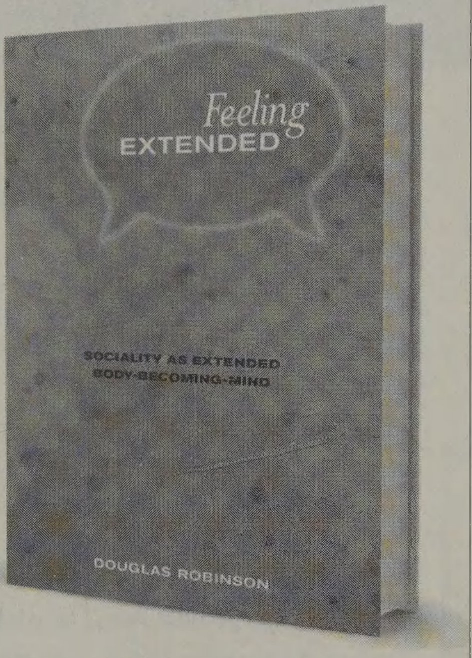
Maimonides
Moshe Halbertal


PETER SEDGWICK
NIETZSCHE'S JUSTICE


THE VITALITY OF CONTRADICTION
BRUCE GILBERT

TRUTH
Matters


THE OUTER LIMITS OF REASON
WHAT SCIENCE, MATHEMATICS, AND LOGIC CANNOT TELL US
NOSON S. YANOFSKY


Feeling EXTENDED
SOCIALITY AS EXTENDED BODY-BECOMING-MIND
DOUGLAS ROBINSON

Tikkun
JUSTICE AS CARING
GET MONEY OUT OF POLITICS
Tikkun
A JUBILEE TO ABOLISH DEBT
A SPIRITUAL WAY OF SEEING
Tikkun
WHAT IS SPIRITUALITY?
AWAY WITH ALL BORDERS
EMBRACING IMMIGRATION AND ENDING DEPORTATION

Submissions

Send articles by e-mail to phil-review@cornell.edu. In the e-mail, include the title of the article, the author's name and postal address, and the author's institutional affiliation. If submitting a paper manuscript, send only one copy to *Philosophical Review*, Cornell University, 217 Goldwin Smith, Ithaca, NY 14853-4601. The manuscript will not be returned unless accompanied by a postage-paid envelope.

Manuscripts must not have been published previously, in part or in whole, or be under consideration for publication elsewhere. Neither the author's name nor the author's affiliation should be included in the manuscript, and any acknowledgments or references to the author's own work should not reveal the author's identity. Any prefatory or explanatory remarks intended for the editors or readers must be given anonymously, not in the cover letter. Manuscripts submitted by members of the Cornell department are accepted only on the recommendation of an outside referee.

Double-space all text, including extracts and notes; all pages must be numbered and have margins of at least one inch on all sides. Manuscripts chosen for publication must eventually conform to *The Chicago Manual of Style*, 16th ed., but for the initial review any clear and consistent citation style is acceptable.

Book Reviews

Unsolicited book reviews are not accepted. Send books for review to Book Review Editor, *Philosophical Review*, Cornell University, 217 Goldwin Smith, Ithaca, NY 14853-4601.